



सत्यमेव जयते

INDIAN AGRICULTURAL
RESEARCH INSTITUTE, NEW DELHI

16199
~::~~

L.A.R I.6.

GIP NLK—H 3 I.A.R.I. --10-5 55--15,000

1942

ANNALS OF MATHEMATICS

(FOUNDED BY ORMOND STONE)

EDITED BY

SOLOMON LEFSCHETZ

JOHN VON NEUMANN

FRÉDÉRIC BOHNENBLUST

WITH THE COÖPERATION OF THE

DEPARTMENT OF MATHEMATICS OF PRINCETON UNIVERSITY

AND

THE SCHOOL OF MATHEMATICS OF THE INSTITUTE
FOR ADVANCED STUDY

AND

A. A. ALBERT	T. H. HILDEBRANDT ²	M. H. STONE
EMIL ARTIN	NATHAN JACOBSON	ANNA PELL-WHEELER ¹
GARRETT BIRKHOFF	SAUNDERS MACLANE ³	G. T. WHYBURN ¹
M. R. HESTENES ¹	N. E. STEENROD	OSCAR ZARISKI

(¹ Representing the MATHEMATICAL ASSOCIATION OF AMERICA)

(² Representing the AMERICAN MATHEMATICAL SOCIETY)

SECOND SERIES, VOL. 43

PUBLISHED QUARTERLY

AT

MOUNT ROYAL AND GUILFORD AVENUES

BALTIMORE, MD.

BY THE

PRINCETON UNIVERSITY PRESS

PRINCETON, N. J.

COPYRIGHT, 1942, BY PRINCETON UNIVERSITY PRESS

A considerable part of the developments in this work refers to equations containing a parameter λ , entering linearly in C . The results for λ non real and F self adjoint are partially analogous to those found in (C). The theory for λ real presents additional complications and has some analogy with a work of W. J. Trjitzinsky⁴ in the field of singular integral equations.

In sections 2, 3 we carry out the transformation into an integral equation (Theorem 3.1). In section 4 the kernel of this equation is investigated in some detail (Theorem 4.1) and the equation is iterated a finite number of times, which yields an integral equation satisfying certain regularity properties. This equation plays an essential role in the subsequent developments. In section 5 we introduce characteristic functions and values (cf. (5.6), (5.9)) for certain approximating homogeneous boundary value problems ((5.7), (5.10)); we establish for these functions orthogonality properties (Theorem 5.1).

The developments in sections 6, 7, 8, 9, 10 are for F self adjoint.

In section 6 we first establish existence of solutions for the non homogeneous problem (5.1) when $\Im \lambda \neq 0$ (Theorem 6.1). In the complex λ -plane we define sets T (Definition 6.1) and particular sets T , which we term O (see text after (6.14)); sets T contain all the points off the axis of reals and may contain certain points on the axis of reals, depending on sufficient 'rarefaction' of the set of points represented by the totality of characteristic values formed for an infinite sequence of approximating homogeneous boundary value problems. We establish existence of solutions, $\subset L_2$, of (5.1) when λ is in T (Theorem 6.2). In section 7 there is developed a corresponding spectral theory (Theorem 7.1). In sections 8 and 9 this theory is applied; we introduce the notion of 'closure' of F with respect to a spectrum θ (Definition 8.1) and give a generalized Fourier expansion of functions, $\subset L_2$ (in D), in terms of θ (Theorem 8.1). Further, there is given a representation for solutions, referred to in Theorem 6.2, with the aid of θ (Theorem 8.2). In section 9 various representations, convergent in the mean square, are given. In section 10 we introduce certain operators $A_\pm(\lambda | \dots)$, with the aid of which certain particular solutions of (5.1) may be expressed for λ in O (Theorem 10.1). With the aid of these operators we investigate conditions under which F is of class I (Theorem 10.2, Definition 6.2) and, finally, questions of uniqueness are studied for λ non real (Theorem 10.3) and for λ in O (Theorem 10.4).

In section 5 a certain "regularity" hypothesis is introduced for the domain D ; this is needed only in the developments subsequent to (5.7). Some lightening of the conditions imposed on the coefficients in (1.1) is possible without any essential change in the conclusions; for simplicity this lightening of the hypothesis will be avoided.

The case when F is of class I corresponds (in the field of partial differential equations), as pointed out by Carleman, to the case of "Grenzpunkt" in the well known theory of H. Weyl in the field of ordinary differential equations.

⁴ W. J. Trjitzinsky, *Some problems in the theory of singular integral equations* [Annals of Math. (1940), 584-619].

Every solution u of (3.37), such that the $D_2 u$ are continuous in D , is solution of (1.1) in the ordinary sense; (3.37) thus generalizes the operator F ; with F assigned this generalized meaning, the partial differential equation and the integral equation are equivalent.

2. Preliminaries

The adjoint of $F(u)$ is

$$(2.1) \quad G(v) = \sum_{i,k} \frac{\partial^2}{\partial x_i \partial x_k} (A_{ik} v) - \sum_i \frac{\partial}{\partial x_i} (B_i v) + C v.$$

The characteristic form for (1.1) is

$$(2.2) \quad A(\gamma_1, \dots, \gamma_m) = \sum_{i,k} A_{i,k} \gamma_i \gamma_k = A(\gamma_1, \dots, \gamma_m; x_1, \dots, x_m).$$

We shall make use of the symbol

$$\frac{\partial}{\partial \nu}$$

denoting operation of derivation along the *transversal direction* (cf. (H; 87)). This operation is to be understood as follows. We form

$$(2.3) \quad A = A(\pi_1, \dots, \pi_m),$$

where the π_i are direction cosines of normal to a surface, say

$$H \equiv H(x_1, \dots, x_m) = 0;$$

that is

$$(2.4) \quad \pi_i = \pm \left[\left(\frac{\partial H}{\partial x_1} \right)^2 + \dots + \left(\frac{\partial H}{\partial x_m} \right)^2 \right]^{-1/2} \frac{\partial H}{\partial x_i},$$

the supposition being, of course, that the $D_1 H$ exist in the domain under consideration. The transversal direction is defined by the ordinary differential system (generally non linear):

$$(2.5) \quad \frac{dx_1}{w_1} = \frac{dx_2}{w_2} = \dots = \frac{dx_m}{w_m} = d\nu,$$

where

$$(2.5a) \quad w_k = w_k(x_1, \dots, x_m) = \sum_{i=1}^m A_{i,k} \pi_i \quad (k = 1, \dots, m).^5$$

In the case when

$$\sum_{i,k} A_{i,k} \frac{\partial^2}{\partial x_i \partial x_k} = \Delta \quad (\text{Laplacian})$$

⁵ See (H; 88) for a detailed discussion of the transversal direction.

the "transversal" direction will signify "normal" direction. In view of (2.5), (2.5a) we are brought to the definition

$$(2.6) \quad \frac{\partial u}{\partial \nu} = \sum_{k=1}^m \frac{\partial u}{\partial x_k} \frac{dx_k}{d\nu} = \sum_{k=1}^m \frac{1}{2} \frac{\partial A}{\partial \pi_k} \frac{\partial u}{\partial x_k} \quad (\text{cf. (2.3)}).$$

There is on hand a following fundamental formula (see (H))

$$(2.7) \quad \int^{(m)} [vF(u) - uG(v)] dt = - \int^{(m-1)} \left[v \frac{\partial u}{\partial \nu} - u \frac{\partial v}{\partial \nu} + Luv \right] dS.$$

[F from (1, 1); G from (2.1)]. Here the first integral displayed is over the space T bounded by the surface $H = 0$; the integral of the second member is over the surface S constituting the frontier of T ;⁶ moreover,

$$(2.7a) \quad L = \sum_i \pi_i B_i - \sum_{i,k} \pi_i \frac{\partial A_{i,k}}{\partial x_k} \quad (\text{cf. (2.4)})$$

and

$$dT = dx_1 \cdots dx_m, \quad dS = \frac{\partial H}{\partial n} \frac{dT}{dH} = \left[\left(\frac{\partial H}{\partial x_1} \right)^2 + \cdots + \left(\frac{\partial H}{\partial x_m} \right)^2 \right]^{\frac{1}{2}} \frac{dT}{dH}.$$

Of importance for us will be a function of (x) and (z) ,

$$\Gamma(x, z) = \Gamma(x_1, \dots, x_m; z_1, \dots, z_m),$$

satisfying the partial differential equation

$$(2.8) \quad A \left(\frac{\partial \Gamma}{\partial x_1}, \dots, \frac{\partial \Gamma}{\partial x_m}; x_1, \dots, x_m \right) = 4\Gamma;$$

namely, the function for which

$$\Gamma^{\sharp}(x; z) = r(x; z)$$

denotes the "geodesic distance"⁸ between the points (x) and (z) . This distance is formed with respect to the quadratic form

$$(2.9) \quad H = H(\gamma_1, \dots, \gamma_m) = \sum_{i,k} H_{i,k}(x) \gamma_i \gamma_k,$$

"associated" with the quadratic form $A(\gamma_1, \dots, \gamma_m)$ (2.2). In this connection it is to be noted that, if

$$(A_{i,j}) \quad (i, j = 1, \dots, m)$$

denoted a matrix with $A_{i,j}$ in the i -th row and j -th column, then

$$(H_{i,j}) = (A_{i,j})^{-1};$$

that is, the matrix $(H_{i,j})$ is the inverse of the matrix $(A_{i,j})$.

With the aid of the developments given in (H; pp. 120-124) we shall study $\Gamma(x; z)$ in greater detail.

⁶ This is asserted under suitable regularity conditions, in the sequel satisfied.

Consider the differential problem

$$(2.10) \quad \frac{dx_i}{ds} = \frac{1}{2} \frac{\partial A}{\partial p_i}, \quad \frac{dp_i}{ds} = -\frac{1}{2} \frac{\partial A}{\partial x_i} \quad (i = 1, \dots, m),$$

where

$$(2.11) \quad A = \sum_{i,k} A_{i,k}(x) p_i p_k,$$

while the initial conditions are

$$(2.11a) \quad x_i = z_i \quad (\text{for } s = 0), \quad p_i = p_{0,i} \quad (\text{for } s = 0);$$

we write

$$A_0(p_0) \equiv \sum_{i,k} A_{i,k}(z) p_{0,i} p_{0,k} = c.$$

As indicated in (H; pp. 120, 121) the solution of this problem may be written in the form

$$(2.12) \quad \begin{aligned} x_i &= \varphi_i(q_1, \dots, q_m; z_1, \dots, z_m) & (q_i = s p_{0,i}), \\ P_i &= \psi_i(q_1, \dots, q_m; z_1, \dots, z_m) & (P_i = s p_i); \end{aligned}$$

from these relations one obtains

$$z_i = \varphi_i(-P_1, \dots, -P_m; x_1, \dots, x_m), \quad q_i = \psi_i(-P_1, \dots, -P_m; x_1, \dots, x_m).$$

It is observed that (2.10) determines geodesics, with respect to the form H , through the point (z) . Moreover, for solutions of (2.10) one has $A(p, x) = \text{constant}$. On substituting $s = 0$ and using the initial conditions this constant is found to be c .

In consequence of known theorems for ordinary differential equations⁷ we infer the following. If the

$$(2.13) \quad A_{i,k}(x), \quad D_\nu A_{i,k}(x) \quad (\nu = 1, \dots, 4)$$

are continuous for (x) in D , which we henceforth assume, then the elements x_i , P_i of the solution (2.12) are continuous and have continuous partial derivatives of the first and second orders with respect to the q_i and the z_i ; furthermore, the third order derivatives will exist. More generally, if the $D_{N+1} A_{i,k}(x)$ are continuous, the partial derivatives with respect to the q_i and z_i , of order $N - 1$, of the x_i and P_i will be continuous and the derivatives of order N will exist. The above statements refer to local properties—for (z) representing a point in D .

According to (H)

$$(2.14) \quad \Gamma(x, z) \equiv A(q_1, \dots, q_m; z_1, \dots, z_m) = \sum_{i,k} A_{i,k}(z) q_i q_k.$$

⁷ Cotton, *Sur les équations différentielles dépendant de paramètres arbitraires* [Bull. Soc. mat., t. 37 (1909), 204–214]. Also see t. 38 (1918), p. 4.

Here the q_i are to be replaced by the expressions, in terms of the z_j and x_j , obtainable from the first m relations (2.12).

We have

$$(2.15) \quad x_i = x_i(s) = x_i(0) + x_i^{(1)}(0) \frac{s}{1!} + (x_i^{(2)}(0) + \xi_i(s)) \frac{s^2}{2!},$$

where

$$\lim_{s \rightarrow 0} \xi_i(s) = 0 \quad (i = 1, \dots, m).$$

By (2.11a) $x_i(0) = z_i$; on the other hand, in view of (2.10) and (2.11)

$$x_i^{(1)}(0) = \frac{1}{2} \frac{\partial A}{\partial p_i}(s=0) = \sum_k A_{i,k}(x) p_k(s=0) = \sum_k A_{i,k}(z) p_{0,k}$$

so that, by definition of q_i ,

$$(2.16) \quad s x_i^{(1)}(0) = \sum_k A_{i,k}(z) q_k.$$

By (2.10)

$$(2.17) \quad x_i^{(2)}(s) = \frac{d}{ds} \sum_\nu A_{i,\nu}(x) p_\nu = \sum_\nu A_{i,\nu}(x) \frac{dp_\nu}{ds} + \sum_\nu p_\nu \frac{dA_{i,\nu}(x)}{ds}.$$

Now the first sum in the last member of (2.17) is

$$g(s) = -\frac{1}{2} \sum_\nu A_{i,\nu}(x) \frac{\partial A}{\partial x_\nu} = -\frac{1}{2} \sum_{\alpha,\beta} A_{i,\nu}(x) \frac{\partial A_{\alpha,\beta}(x)}{\partial x_\nu} p_\alpha p_\beta$$

moreover, since $q_j = s p_{0,j}$,

$$(2.17a) \quad s^2 g(0) = -\frac{1}{2} \sum_{\alpha,\beta} A_{i,\nu}(z) \frac{\partial A_{\alpha,\beta}(z)}{\partial z_\nu} q_\alpha q_\beta.$$

We have

$$\frac{dA_{i,\nu}(x)}{ds} = \sum_{\gamma=1}^m \frac{\partial A_{i,\nu}(x)}{\partial x_\gamma} \frac{dx_\gamma}{ds};$$

whence by (2.16)

$$\left[\frac{dA_{i,\nu}(x)}{ds} \right]_{s=0} s = \sum_{\gamma=1}^m \frac{\partial A_{i,\nu}(z)}{\partial z_\gamma} s x_\gamma^{(1)}(0) = \sum_{\gamma,k} A_{\gamma,k}(z) \frac{\partial A_{i,\nu}(z)}{\partial z_\gamma} q_k$$

and

$$(2.17b) \quad \left[\frac{1}{2} \sum_\nu p_\nu \frac{dA_{i,\nu}(x)}{ds} \right]_{s=0} s^2 = \sum_{\gamma,k} \frac{1}{2} A_{\gamma,k}(z) \frac{\partial A_{i,\nu}(z)}{\partial z_\gamma} q_k q_\nu.$$

In consequence of (2.17a), (2.17b) from (2.17) one infers that

$$(2.18) \quad \frac{s^2}{2!} x_i^{(2)}(0) = \frac{s^2}{2} g(0) + \text{last member of (2.17b)} = \sum_{\alpha,\beta} A_{i;\alpha,\beta}(z) q_\alpha q_\beta,$$

$$A_{i;\alpha,\beta}(z) = -\frac{1}{4} \sum_\nu A_{i,\nu}(z) \frac{\partial A_{\alpha,\beta}(z)}{\partial z_\nu} + \frac{1}{2} \sum_\nu A_{\nu,\beta}(z) \frac{\partial A_{i,\alpha}(z)}{\partial z_\nu}.$$

Finally, one may write (2.15) in the form

$$(2.19) \quad x_i - z_i = \sum_k A_{i,k}(z)q_k + \sum_{\alpha,\beta} (A_{i;\alpha,\beta}(z) + \lambda_{i;\alpha,\beta}(z; q))q_\alpha q_\beta,$$

where the $A_{i;\alpha,\beta}(z)$ are from (2.18) and

$$\lim \lambda_{i;\alpha,\beta}(z; q) = 0 \quad (\text{as } (q) \rightarrow (0); (z) \text{ in } D);$$

furthermore, the $\lambda_{i;\alpha,\beta}$ are continuous in

$$z_1, \dots, z_m, q_1, \dots, q_m \quad (|q_j| \leq q_0; q_0 > 0; (z) \text{ in } D).$$

Inversion of (2.19) will give

$$(2.20) \quad q_i = \sum_k H_{i,k}(z)(x_k - z_k) + \sum_{\alpha,\beta} (H_{i;\alpha,\beta}(z) + \bar{\lambda}_{i;\alpha,\beta})(x_\alpha - z_\alpha)(x_\beta - z_\beta),$$

where

$$(2.20a) \quad H_{i;\sigma,\tau}(z) = - \sum_{\alpha,\beta,k} A_{k;\alpha,\beta}(z) H_{i,k}(z) H_{\alpha,\sigma}(z) H_{\beta,\tau}(z) \quad (\text{cf. (2.18)})$$

and the $\bar{\lambda}$ are continuous in the x_j, z_j , while

$$(2.20b) \quad \lim \bar{\lambda}_{i;\alpha,\beta} = \lim \bar{\lambda}_{i;\alpha,\beta}(z; x) = 0 \quad (\text{as } (x) \rightarrow (z); (z) \text{ in } D).$$

We note that the $H_{i;\sigma,\tau}$ are polynomials in the

$$A_{i,j}(z), \quad D_1 A_{i,j}(z), \quad H_{i,j}(z).$$

Accordingly, the functions of (2.20a) are continuous in D .

Substitution of (2.20) into (2.14) will yield

$$\begin{aligned} \Gamma(x, z) = & \sum_{k,k_1} H_{k,k_1}(z)(x_k - z_k)(x_{k_1} - z_{k_1}) \\ & + 2 \sum_{\alpha,\beta,k_1} \bar{H}_{k_1;\alpha,\beta}(x, z)(x_\alpha - z_\alpha)(x_\beta - z_\beta)(x_{k_1} - z_{k_1}) + \Gamma_4, \end{aligned}$$

where

$$\bar{H}_{k_1;\alpha,\beta}(x, z) = H_{k_1;\alpha,\beta}(z) + \bar{\lambda}_{k_1;\alpha,\beta} \quad (\text{cf. (2.20a), (2.20b)})$$

and

$$\Gamma_4 = \sum_{\alpha,\beta,\alpha_1,\beta_1} (x_\alpha - z_\alpha)(x_\beta - z_\beta)(x_{\alpha_1} - z_{\alpha_1})(x_{\beta_1} - z_{\beta_1}) \sum_{r,\nu} A_{r,\nu}(z) \bar{H}_{r;\alpha,\beta} \bar{H}_{\nu;\alpha_1,\beta_1}.$$

Finally, we obtain

$$(2.21) \quad \Gamma(x, z) = \sum_{k,k_1} H_{k,k_1}(z)(x_k - z_k)(x_{k_1} - z_{k_1}) + \Gamma_3(x, z),$$

where

$$(2.21a) \quad \Gamma_3(x, z) \equiv \sum_{\alpha,\beta,k_1} (2H_{k_1;\alpha,\beta}(z) + \lambda_{k_1;\alpha,\beta}^*(x, z))(x_\alpha - z_\alpha)(x_\beta - z_\beta)(x_{k_1} - z_{k_1})$$

(cf. (2.20a)); here

$$(2.21b) \quad \lambda_{k_1;\alpha,\beta}^*(x, z) \rightarrow 0 \quad (\text{as } (x) \rightarrow (z); (z) \text{ in } D);$$

the functions, involved in (2.21b), are continuous in (x) for (x) in a neighborhood of (z) .

When

$$\sum_{i,k} A_{i,k} \frac{\partial^2}{\partial x_i \partial x_k} = \Delta \quad (\text{Laplacian}),$$

we have

$$\Gamma(x, z) = (x_1 - z_1)^2 + \dots + (x_m - z_m)^2.$$

3. The transformation

In view of the definition (2.6) one has

$$(3.1) \quad \frac{\partial}{\partial \nu} \Gamma(x, z) = \sum_{k=1}^m \frac{1}{2} \frac{\partial A}{\partial \pi_k} \frac{\partial \Gamma}{\partial x_k} \quad (\text{cf. (2.4), (2.3)}),$$

the derivative in the transversal direction being with respect to a surface $H = 0$. With $\rho(> 0)$ suitably small we consider the surface

$$(3.2) \quad H_\rho(x) \equiv \Gamma(x, z) - \rho^2 = 0 \quad ((z) \text{ fixed in } D).$$

Correspondingly

$$(3.3) \quad \pi_i = \frac{\partial \Gamma}{\partial x_i} \frac{1}{\sigma(x, z)}, \quad \sigma^2(x, z) = \left(\frac{\partial \Gamma}{\partial x_1} \right)^2 + \dots + \left(\frac{\partial \Gamma}{\partial x_m} \right)^2.$$

By (2.3) and (3.1)

$$\frac{\partial}{\partial \nu} \Gamma(x, z) = \sum_{k=1}^m \sum_{\alpha} A_{\alpha,k}(x) \pi_{\alpha} \frac{\partial \Gamma}{\partial x_k}.$$

Thus, in view of (3.3),

$$\frac{\partial}{\partial \nu} \Gamma(x, z) = \sum_{\alpha,k} A_{\alpha,k}(x) \frac{\partial \Gamma}{\partial x_{\alpha}} \frac{\partial \Gamma}{\partial x_k} \frac{1}{\sigma(x, z)}$$

and, by virtue of (2.8),

$$\frac{\partial}{\partial \nu} \Gamma(x, z) = \frac{1}{\sigma(x, z)} 4\Gamma(x, z).$$

Whence, when α is a constant,

$$(3.4) \quad \frac{\partial}{\partial \nu} \Gamma^{\alpha}(x, z) = \alpha \Gamma^{\alpha-1}(x, z) \frac{\partial}{\partial \nu} \Gamma(x, z) = \frac{4\alpha}{\sigma(x, z)} \Gamma^{\alpha}(x, z) \quad (\text{cf. (3.3)}).$$

With a view to applying the fundamental formula (2.7) we wish to find a function

$$(3.5) \quad v(x, z) = T(\Gamma(x, z))$$

such that

$$(3.5a) \quad v = \frac{\partial v}{\partial \nu} = 0 \quad ((z) \text{ fixed in } D)$$

for (x) on the surface $H_\rho(x) = 0$ (3.2).

On taking account of (3.4) it is observed that a function (3.5) satisfying (3.5a) may be taken of the form

$$(3.6) \quad v(x, z) = \Gamma^{1(2-m)}(x, z) + \frac{1}{\rho^{2m-4}} \Gamma^{1(m-2)}(x, z) - \frac{2}{\rho^{m-2}}.$$

Inasmuch as $m > 2$, it is noted that v becomes infinite essentially as Γ^{1-m} , when $(x) \rightarrow (z)$.

With

$$(3.7) \quad r^2 = (x_1 - z_1)^2 + \cdots + (x_m - z_m)^2 \quad ((z) \text{ fixed in } D)$$

the differential element of volume, dT , may be given in the form

$$(3.7a) \quad dT = r^{m-1} dr d\Omega,$$

where $d\Omega$ is element of surface of a hypersphere of radius unity in the m -space of (x) . We recall that the total surface of this hypersphere is $S_m = 2\pi^{1/2} / g\left(\frac{m}{2}\right)$,

where $g(u)$ is the Gamma-function. The volume is $\frac{1}{m} S_m$. From (3.7a) it is inferred that the element of surface of a hypersphere of radius r is

$$(3.7b) \quad dw = \frac{dT}{dr} = r^{m-1} d\Omega.$$

The element dS of the surface (3.2) satisfies the relation

$$(3.8) \quad dS \cos \tau = dw,$$

where τ is the angle formed by the line extending from (z) and having direction cosines $(x_i - z_i)/r$ and by a suitable direction of the normal to the surface (3.2) at (x) ; the direction cosines of this normal are the π_i of (3.3) (with a suitable sign for $\sigma(x, z)$); thus,

$$\cos \tau = \frac{1}{r\sigma(x, z)} \sum_i \frac{\partial \Gamma(x, z)}{\partial x_i} (x_i - z_i)$$

and, by (3.8) and (3.7b),

$$(3.9) \quad dS = \frac{\sigma(x, z) r^m d\Omega}{\sum_i (x_i - z_i) \frac{\partial \Gamma(x, z)}{\partial x_i}} \quad (\text{cf. (3.3)}).$$

Let K_δ denote a sphere with center at (z) and radius δ ($0 < \delta \leq \delta_0$), where δ_0 is taken sufficiently small so that the surface S_δ of K_δ lies interior the domain Γ_ρ bounded by the surface (3.2).⁸ We have a domain

$$(3.10) \quad D(\rho, \delta),$$

whose frontier consists of the spherical surface S_δ and of the surface $H_\rho(x) = 0$.

⁸ Γ_ρ contains (z) in the interior and is topologically a sphere.

Suppose $u = u(x)$ is a function, continuous with the D_1u , satisfying in D the differential equation (1.1). We substitute in the fundamental formula (2.7) u equal to the aforesaid function and $v = v(x, z)$, from (3.6), and we extend the integrations over the domain $D(\rho, \delta)$ (cf. (3.10)) and over the frontier of $D(\rho, \delta)$; the variable of integration is to be (x) . Thus,

$$(3.10a) \quad \int_{D(\rho, \delta)}^{(m)} [vf - uG(v)] dT = Q(\rho, \delta)$$

where

$$Q(\rho, \delta) = - \int_{H_\rho + S_\delta}^{(m-1)} \left[v \frac{\partial u}{\partial \nu} - u \frac{\partial v}{\partial \nu} + Luw \right] dS.$$

On taking account of (3.5a) one obtains

$$(3.11) \quad Q(\rho, \delta) = - \int_{S_\delta}^{(m-1)} \left[v \frac{\partial u}{\partial \nu} - u \frac{\partial v}{\partial \nu} + Luw \right] dS.$$

With the notation (3.7) we have

$$r = \delta \quad (\text{for } (x) \text{ on } S_\delta);$$

moreover, by virtue of the statement referring to (3.7b), dS in (3.11) is of the form

$$(3.11a) \quad dS = dw(r = \delta) = \delta^{m-1} d\Omega.$$

In (3.11) L is defined by (2.7a), where the π_i are the direction cosines of the normal to the surface S_δ , extended in the direction interior $D(\rho, \delta)$; that is, the normal in question is directed outwards from the sphere K_δ . We have

$$\pi_i = \frac{x_i - z_i}{\delta} \quad ((x) \text{ in } S_\delta)$$

and

$$(3.11b) \quad L = \sum_i B_i(x) \pi_i - \sum_{i,k} \frac{\partial A_{i,k}(x)}{\partial x_k} \pi_i.$$

By definition (2.6)

$$(3.21) \quad \frac{\partial u}{\partial \nu} = \sum_{k=1}^m \frac{1}{2} \frac{\partial A}{\partial \pi_k} \frac{\partial u}{\partial x_k} = \sum_{k,i} A_{i,k}(x) \pi_i \frac{\partial u}{\partial x_k} \quad (\text{on } S_\delta).$$

Now

$$(3.13) \quad \frac{\partial \Gamma}{\partial \nu} = \sum_{k,i} A_{i,k}(x) \pi_i \frac{\partial \Gamma}{\partial x_k}$$

and, in consequence of (3.6),

$$(3.14) \quad \frac{\partial v}{\partial \nu} = \left[\frac{2-m}{2} \Gamma^{-1m}(x, z) + \frac{1}{\rho^{2m-4}} \left(\frac{m-2}{2} \right) \Gamma^{1m-2}(x, z) \right] \frac{\partial \Gamma}{\partial \nu}.$$

Before proceeding further we shall study the function

$$(3.15) \quad w(x, z) = \frac{r}{\sqrt{\Gamma(x, z)}} \quad (r^2 = (x_1 - z_1)^2 + \dots + (x_m - z_m)^2).$$

By (2.21)

$$(3.16) \quad w^2(x, z) = [H(z, \pi) + g_s(x, z)]^{-1}$$

where

$$(3.16a) \quad H(z, \pi) = \sum_{k, k_1} H_{k, k_1}(z) \pi_k \pi_{k_1}, \quad \pi_k = \frac{x_k - z_k}{r}$$

and

$$(3.16b) \quad g_s(x, z) = \frac{1}{r^2} \Gamma_s(x, z) \quad (\text{cf. (2.21a)}).$$

For $r \leq \delta_0$

$$(3.16c) \quad |2H_{k_1; \alpha, \beta}(z) + \lambda_{k_1; \alpha, \beta}^*(x, z)| \leq \lambda_0(z),$$

where $\lambda_0(z)$ is independent of (x) . Accordingly, by (3.16b)

$$(3.16d) \quad |g_s(x, z)| \leq r \sum_{\alpha, \beta, k_1} \lambda_0(z) \pi_\alpha \pi_\beta \pi_{k_1} \leq r \lambda(z) \quad (0 \leq r \leq \delta_0)$$

($\lambda(z)$ independent of (x)). Now $H(z, \pi)$ is a positive definite quadratic form; thus, inasmuch as $\pi_1^2 + \dots + \pi_m^2 = 1$, one has

$$(3.17) \quad H(z, \pi) \geq h(z) = \sum_{k, k_1} H_{k, k_1}(z) \pi'_k \pi'_{k_1} > 0$$

(for some values π'_k , with $\pi_1'^2 + \dots + \pi_m'^2 = 1$), where $h(z)$ is independent of (π) . Hence, in view of (3.17) and (3.16d)

$$(3.17a) \quad H(z, \pi) + g_s(x, z) \geq h(z) - r \lambda(z) \geq h(z) - \delta_0 \lambda(z) = h_0^2(z) > 0$$

for $r \leq \delta_0$, provided one takes

$$(3.17b) \quad 0 < \delta_0 < \frac{h(z)}{\lambda(z)}.$$

Whence, by (3.16)

$$(3.18) \quad |w(x, z)| \leq \frac{1}{h_0(z)} < \infty \quad (0 \leq r \leq \delta_0; \text{cf. (3.15), (3.17a), (3.17b)}).$$

It is essential to note that in (3.18) $h_0(z)$ is independent of (x) .

On writing

$$w^2(x, z) = \frac{1}{H(z, \pi)} + \rho_0(x, z),$$

by (3.16), (3.16d), (3.17) and (3.17a) we have

$$|\rho_0(x, z)| = \frac{|g_s(x, z)|}{|H(z, \pi)| |H(z, \pi) + g_s(x, z)|} \leq \frac{r\lambda(z)}{h(z)h_0^2(z)}$$

for $0 \leq r \leq \delta_0$. Accordingly, the function of (3.15) satisfies the relation

$$(3.19) \quad w^2(x, z) = \frac{1}{H(z, \pi)} + r\rho(x, z) \quad (\text{cf. (3.16a)})$$

where

$$(3.19a) \quad |\rho(x, z)| \leq \rho(z) = \frac{\lambda(z)}{h(z)h_0^2(z)} \quad (r \leq \delta_0);$$

here $\rho(z)$ is independent of (x) . In particular

$$(3.19b) \quad \lim_{r \rightarrow 0} w(x, z) = H^{-1}(z, \pi),$$

provided $(x) \rightarrow (z)$ along a fixed direction whose direction cosines are π_1, \dots, π_m . In (3.19b) the limit is attained uniformly with respect to the π_i ; that is, with respect to direction of approach. Of course, the limit depends on the direction of approach, but there is uniformity in the sense implied by (3.19), (3.19a).

The function of (3.11) will be expressed in the form

$$(3.20) \quad Q(\rho, \delta) = Q_1(\rho, \delta) + Q_2(\rho, \delta),$$

where, on taking account of (3.11a),

$$(3.20a) \quad Q_1(\rho, \delta) = - \int^{(m-1)} \left[\frac{\partial u}{\partial \nu} + Lu \right] \nu \delta^{m-1} d\Omega \quad (\text{cf. (3.11b), (3.12)}),$$

and

$$(3.20b) \quad Q_2(\rho, \delta) = \int^{(m-1)} u \frac{\partial \nu}{\partial \nu} \delta^{m-1} d\Omega \quad (\text{cf. (3.14), (3.13)}).$$

If we denote by l the upper bound in the spherical domain K_δ of

$$\sum_i |B_i(x)| + \sum_{i,k} \left| \frac{\partial A_{i,k}(x)}{\partial x_k} \right|,$$

in consequence of (3.11b) it is deduced that

$$|L| \leq l \quad (\text{on } S_\delta; \text{ for } 0 < \delta \leq \delta_0).$$

On the other hand, by (3.12), for (x) on S_δ one has

$$(3.21) \quad \left| \frac{\partial u}{\partial \nu} \right| \leq \sum_{k,i} |A_{i,k}(x)| \left| \frac{\partial u}{\partial x_k} \right| \leq l_1 \quad (0 < \delta \leq \delta_0),$$

where l_1 is independent of δ . In this connection one may define l_1 as the upper-bound of the second member in (3.21) for (x) in the spherical domain K_δ .⁹ Similarly,

$$|u| \leq l_2 \quad (\text{on } S_\delta; l_2 \text{ independent of } \delta (\leq \delta_0)).$$

In consequence of (3.6)

$$(3.22) \quad v\delta^{m-1} = \delta \left\{ \left(\frac{\delta}{\sqrt{\Gamma}} \right)^{m-2} + f(x, z)\delta^{m-2} \right\} \quad ((x) \text{ on } S_\delta),$$

where

$$f(x, z) = \frac{1}{\rho^{2m-4}} \Gamma^{(m-2)}(x, z) - \frac{2}{\rho^{m-2}}.$$

Clearly

$$f(x, z) \leq f_0(z) < \infty \quad (0 \leq r \leq \delta_0),$$

$f_0(z)$ being independent of (x) . Thus, by (3.22) and (3.18)

$$(3.23) \quad |v\delta^{m-1}| \leq \delta \{h_0^{-m+2}(z) + f_0(z)\delta^{m-2}\} \leq \delta f_1(z) \quad ((x) \text{ on } S_\delta; 0 < \delta \leq \delta_0)$$

with $f_1(z) (< \infty)$ independent of (x) and δ .

On taking account of (3.21) and (3.23), from (3.20a) it is inferred that

$$|Q_1(\rho, \delta)| \leq \delta \int^{(m-1)} (l_1 + l_2) f_1(z) d\Omega = \delta(l_1 + l_2) f_1(z) S_m;$$

here S_m is the area of the surface of a hypersphere (in the space of (x)) of radius unity. Consequently

$$(3.24) \quad \lim_{\delta \rightarrow 0} Q_1(\rho, \delta) = 0.$$

Turning our attention to the integral (3.20b) we investigate the integrand involved. In view of (2.21) and of certain other considerations of section 2 one has

$$\frac{\partial \Gamma}{\partial x_k} = \sum_{k_1} (2H_{k,k_1}(x) + \omega_{k,k_1}(z, x))(x_{k_1} - z_{k_1}),$$

where the ω are continuous in (z) and (x) , while

$$(3.25) \quad \omega_{k,k_1}(z, x) \rightarrow 0 \quad (\text{as } (x) \rightarrow (z))$$

uniformly with respect to direction of approach. Whence by virtue of (3.13)

$$\begin{aligned} \frac{\partial \Gamma}{\partial \nu} &= \sum_{k,i} A_{i,k}(x) \pi_i \sum_{k_1} (2H_{k,k_1}(x) + \omega_{k,k_1}(z, x))(x_{k_1} - z_{k_1}) \\ &= \sum_{k,i,k_1} A_{i,k}(x) \omega_{k,k_1}(z, x) \pi_i (x_{k_1} - z_{k_1}) + \gamma, \end{aligned}$$

⁹ It is implied that the $\frac{\partial u}{\partial x_k}$ are continuous in D .

where

$$\gamma = 2 \sum_{i,k_1} \pi_i(x_{k_1} - z_{k_1}) \sum_k A_{i,k}(x) H_{k,k_1}(x) = 2 \sum_{i=1}^m \pi_i(x_i - z_i) = 2\delta.$$

Hence, the transversal derivative being with respect to the spherical surface S_i ,

$$\frac{1}{\delta} \frac{\partial \Gamma}{\partial \nu} = 2 + \omega'(z, x) \quad ((x) \text{ on } S_i);$$

here $\omega'(z, x)$ is continuous for (x) in Γ_ρ and (in view of (3.25))

$$(3.26) \quad |\omega'(z, x)| \leq \sum_{k,i,k_1} |A_{i,k}(x)| |\omega_{k,k_1}(z, x)| \rightarrow 0$$

(as $(x) \rightarrow (z)$) *uniformly* with respect to the direction along which (x) approaches the fixed point (z) in D . Whence, on taking account of (3.15), it is inferred that

$$(3.27) \quad \Gamma^{-1} \frac{\partial \Gamma}{\partial \nu} = w(x, z)(2 + \omega'(z, x)) \quad ((x) \text{ on } S_i).$$

By virtue of (3.14)

$$\frac{\partial v}{\partial \nu} \delta^{m-1} = \left[\frac{2-m}{2} (w(x, z))^{m-1} + \frac{1}{\rho^{2m-4}} \left(\frac{m-2}{2} \right) \Gamma^{1+m-1} \delta^{m-1} \right] \Gamma^{-1} \frac{\partial \Gamma}{\partial \nu}.$$

Since $m \geq 3$ we have

$$(3.28) \quad \frac{\partial v}{\partial \nu} \delta^{m-1} = \left[\frac{2-m}{2} w^{m-1}(x, z) + \delta^{m-1} w_1(x, z) \right] \Gamma^{-1} \frac{\partial \Gamma}{\partial \nu},$$

where

$$(3.29) \quad |w_1(x, z)| = \left| \frac{1}{\rho^{2m-4}} \left(\frac{m-2}{2} \right) (\sqrt{\Gamma})^{m-3} \right| \leq w_1(z) < \infty,$$

$w_1(z)$ being independent of (x) . Finally, substitution of (3.27) in (3.28) will yield

$$(3.30) \quad \frac{\partial v}{\partial \nu} \delta^{m-1} = (2-m)w^m(x, z) + R(x, z),$$

with

$$R(x, z) = 2\delta^{m-1} w_1(x, z) w(x, z) + \omega'(z, x) \left[\frac{2-m}{2} w^m(x, z) + \delta^{m-1} w_1(x, z) w(x, z) \right].$$

In view of (3.29), (3.18) and of (3.26) it is observed that $R(x, z)$ is *continuous in* (x) , for (x) in Γ_ρ ; moreover,

$$R(x, z) \rightarrow 0 \quad (\text{as } (x) \rightarrow (z)),$$

uniformly with respect to the direction along which $(x) \rightarrow (z)$.

On taking account of continuity of u , in consequence of (3.30) we have

$$(3.31) \quad u(x) \frac{\partial v}{\partial \nu} \delta^{m-1} = (u(z) + u(z, x))[(2 - m)w^m(x, z) + R(x, z)]$$

where $u(z, x)$ is continuous in \bar{z} , (x) ((z) , (x) in D) and $\rightarrow 0$ uniformly (for (z) fixed in D), as $(x) \rightarrow (z)$.¹⁰ Relation (3.31), together with (3.19b) (satisfied uniformly in the previously indicated sense), leads to the conclusion that for the integral (3.20b) one has

$$(3.32) \quad \lim_{\delta \rightarrow 0} Q_2(\rho, \delta) = \int^{(m-1)} \frac{u(z)(2 - m)}{(\sqrt{H(z, \pi)})^m} d\Omega.$$

Thus, by (3.20), (3.32) and (3.24), *there is on hand the relation*

$$(3.33) \quad \lim_{\delta \rightarrow 0} Q(\rho, \delta) = -k(z)u(z),$$

where

$$(3.33a) \quad k(z) = (m - 2) \int^{(m-1)} \frac{d\Omega}{\sqrt{(H(z, \pi))^m}},$$

with

$$H(z, \pi) = \sum_{i,j} H_{i,j}(z) \pi_i \pi_j.$$

In (3.33a) $d\Omega$ is the differential element of surface of a hypersphere of radius unity, $(\pi) = (\pi_1, \dots, \pi_m)$ is thought of as representing the variable point (on this sphere) with respect to which the integration is performed.

At this stage it will be convenient to obtain certain inequalities for $k(z)$. By (3.33a) and (3.17)

$$(3.34) \quad k(z) \leq (m - 2)h^{-\frac{1}{2}m}(z) \int^{(m-1)} d\Omega = (m - 2)h^{-\frac{1}{2}m}(z)S_m.$$

On the other hand,

$$(3.35) \quad H(z, \pi) \leq H(z) < \infty \quad (\text{for } (\pi) \text{ on unit sphere})$$

for (z) in D ; here $H(z)$ may be taken continuous (independent of (π)). Accordingly

$$(3.36) \quad k(z) \geq (m - 2) \int^{(m-1)} H^{-\frac{1}{2}m}(z) d\Omega = (m - 2)H^{-\frac{1}{2}m}(z)S_m > 0.$$

By virtue of (3.10a) and (3.33) it is concluded that

$$-k(z)u(z) = \int_{\Gamma_p}^{(m)} [v(x, z)f(x) - u(x)G(v(x, z))] dT_x,$$

¹⁰ Uniformly with respect to direction of approach.

where Γ_ρ is the domain, containing (z) in the interior, bounded by the surface $\Gamma(x, z) - \rho^2 = 0$ ($\rho(> 0)$ sufficiently small). Accordingly, u satisfies the integral equation

$$(3.37) \quad u(z) - \int_{\Gamma}^{(m)} K(x, z) u(x) dT_x = f_o(z),$$

with

$$(3.37a) \quad f_o(z) = - \int_{\Gamma_o}^{(m)} \frac{v(x, z)}{k(z)} f(x) dT_x,$$

$$(3.37b) \quad K(x, z) = \frac{1}{k(z)} G(v(x, z)).$$

We sum the above developments as follows.

THEOREM 3.1. *Every solution u of (1.1) which, together with the $D_1 u$ is continuous in the domain D , satisfies the integral equation (3.37). Here $f_o(z)$ is given by (3.37a); the kernel $K(x, z)$ is defined by (3.37b). The domain Γ_ρ contains (z) and is bounded by the surface $\Gamma(x, z) - \rho^2 = 0$ [$\rho(> 0)$, sufficiently small; (z) in D]. The function $v(x, z)$ involved in (3.37a) and (3.37b) is of the form*

$$v(x, z) = \Gamma^{1(2-m)}(x, z) + \frac{1}{\rho^{2m-4}} \Gamma^{1(m-2)}(x, z) - \frac{2}{\rho^{m-2}} = v(z, x).$$

The differential operator G is the adjoint of F and is, accordingly, defined by (2.1) (the derivatives are with respect to the x_j). Finally, $k(z)$ is the function of (3.33a) and satisfies the inequalities (3.34) and (3.36) for (z) in D .

We define $v(x, z)$ and, hence, $K(x, z)$ as zero for (x) (in D) exterior Γ_ρ (i.e. for (x) such that $\Gamma(x, z) \geq \rho^2$, when $\rho(> 0)$ possibly depends on (z)). Then $K(x, z)$ will be defined for all (x) and (z) in D . With this in view, the field of integration in (3.37), (3.37a) may be replaced by D .

4. Investigation of the integral equation

In consequence of the developments of section 2 it is inferred that

$$(4.1) \quad \frac{\partial \Gamma(x, z)}{\partial x_k} = \sum_{k_1} (2H_{k, k_1}(x) + o_{k, k_1}(z, x))(x_{k_1} - z_{k_1}),$$

$$(4.1a) \quad \frac{\partial^2 \Gamma}{\partial x_k \partial x_{k_1}} = 2H_{k, k_1}(x) + o'_{k, k_1}(z, x),$$

where the functions

$$o_{k, k_1}(z, x), \quad o'_{k, k_1}(z, x)$$

are continuous in (x) for (x) in Γ_ρ , while

$$o_{k, k_1}(z, x), \quad o'_{k, k_1}(z, x) \rightarrow 0 \quad (\text{as } (x) \rightarrow (z)).$$

Now

$$\sum_{i, k} \frac{\partial^2}{\partial x_i \partial x_k} (A_{i, k} v) = \sum_{i, k} A_{i, k} \frac{\partial^2 v}{\partial x_i \partial x_k} + 2 \sum_{i, k} \frac{\partial A_{i, k}}{\partial x_i} \frac{\partial v}{\partial x_k} + v \sum_{i, k} \frac{\partial^2}{\partial x_i \partial x_k} A_{i, k};$$

hence by (2.1)

$$(4.2) \quad G(v) = \sum_{i,k} A_{i,k}(x) \frac{\partial^2 v}{\partial x_i \partial x_k} + \sum_k B'_k(x) \frac{\partial v}{\partial x_k} + C'(x)v,$$

where

$$(4.3) \quad \begin{aligned} B'_k(x) &= 2 \sum_i \frac{\partial A_{i,k}}{\partial x_i} - B_k, \\ C'(x) &= \sum_{i,k} \frac{\partial^2}{\partial x_i \partial x_k} A_{i,k} - \sum_i \frac{\partial B_i}{\partial x_i} + C. \end{aligned}$$

By (3.6)

$$v = v(x, z) = T(\Gamma(x, z)), \quad T(\zeta) = \zeta^{\frac{1}{2}(2-m)} + \frac{1}{\rho^{2m-4}} \zeta^{\frac{1}{2}(m-2)} - \frac{2}{\rho^{m-2}}.$$

Thus

$$\frac{\partial v}{\partial x_k} = T^{(1)}(\Gamma) \frac{\partial \Gamma}{\partial x_k}, \quad \frac{\partial^2 v}{\partial x_i \partial x_k} = T^{(2)}(\Gamma) \frac{\partial \Gamma}{\partial x_i} \frac{\partial \Gamma}{\partial x_k} + T^{(1)}(\Gamma) \frac{\partial^2 \Gamma}{\partial x_i \partial x_k}$$

and, by virtue of (4.2),

$$\begin{aligned} G(v) &= T^{(1)}(\Gamma) \left\{ \sum_{i,k} A_{i,k}(x) \frac{\partial^2 \Gamma}{\partial x_i \partial x_k} + \sum_k B'_k(x) \frac{\partial \Gamma}{\partial x_k} \right\} \\ &\quad + C'(x)T(\Gamma) + T^{(2)}(\Gamma) \sum_{i,k} A_{i,k}(x) \frac{\partial \Gamma}{\partial x_i} \frac{\partial \Gamma}{\partial x_k}. \end{aligned}$$

On making use of (4.2) and of the differential equation (2.8), we accordingly obtain

$$(4.4) \quad G(v) = \tau(x, z)T^{(1)}(\Gamma) + C^1(x)T(\Gamma) + 4\Gamma T^{(2)}(\Gamma),$$

where

$$(4.4a) \quad \tau(x, z) = \sum_{i,k} A_{i,k}(x) \frac{\partial^2 \Gamma}{\partial x_i \partial x_k} + \sum_k B'_k(x) \frac{\partial \Gamma}{\partial x_k} = G(\Gamma) - C'(x)\Gamma;$$

here

$$(4.4b) \quad \begin{aligned} T^{(1)}(\Gamma) &= \frac{m-2}{2} \left[\frac{1}{\rho^{2m-4}} \Gamma^{\frac{1}{2}(m-4)} - \Gamma^{-\frac{1}{2}m} \right], \\ T^{(2)}(\Gamma) &= \frac{m-2}{2} \left[\frac{1}{\rho^{2m-4}} \left(\frac{m-4}{2} \right) \Gamma^{\frac{1}{2}(m-6)} + \frac{m}{2} \Gamma^{\frac{1}{2}(-m-2)} \right]. \end{aligned}$$

From (4.4a) by virtue of (4.1) and (4.12) we deduce

$$(4.5) \quad \begin{aligned} \tau(x, z) &= 2m + \tau_1(x, z), \\ \tau_1(x, z) &= \sum_{i,k} A_{i,k}(x) o'_{i,k}(z, x) + \sum_{k,k_1} B'_k(x) (2H_{k,k_1}(x) + o_{k,k_1}(z, x))(x_{k_1} - z_{k_1}). \end{aligned}$$

It is observed that $\tau(x, z)$ is independent of $C(x)$; moreover, $\tau(x, z)$ is continuous in (x) for (x) in Γ_ρ . Also

$$(4.5a) \quad \tau_1(x, z) \rightarrow 0 \quad (\text{as } (x) \rightarrow (z)).$$

It is possible to obtain a more detailed expression for $\tau_1(x, z)$, namely

$$(4.6) \quad \tau_1(x, z) = \sum_i (\tau_{1,i}(z) + \delta_i(z, x))(x_i - z_i);$$

here the $\tau_{i,i}(z)$ are polynomials in the

$$H_{i,k}(z), A_{i,k}(z), D_1 A_{i,k}(z), B_i(z);$$

moreover, the $\delta_i(z, x)$ are continuous in (x) for (x) in Γ_ρ , while

$$(4.6a) \quad \delta_i(z, x) \rightarrow 0 \quad (\text{as } (x) \rightarrow (z)).$$

The precise form of the $\tau_{1,i}(z)$ in (4.6) is

$$(4.6b) \quad \tau_{1,i}(z) = \lambda_i^0(z) + \sum_{i,k} 2H_{i,k}(z) \frac{\partial A_{i,k}(z)}{\partial z_j},$$

where

$$(4.6c) \quad \lambda_i^0(z) = \sum_{i,\nu} A_{i,\nu}(z) N_{i,\nu}^i(z) + 2 \sum_i B_i'(z) H_{i,i}(z),$$

with

$$(4.6d) \quad N_{i,\nu}^i(z) = T_{\nu,i}^i + T_{i,\nu}^i + T_{\nu,i}^i + T_{i,\nu}^i + T_{i,i}^i + T_{i,i}^i,$$

$$(4.6e) \quad T_{k,\mu}^i = -\sum_{\tau,\gamma} H_{\tau,\mu} H_{\gamma,i} \frac{\partial A_{\gamma,\tau}}{\partial z_k} + \frac{1}{2} \sum_{i,\tau} H_{i,k} H_{\tau,\mu} \frac{\partial A_{i,\tau}}{\partial z_i}.$$

By (4.4), (4.5) and (4.4b)

$$G(v) = (2m + \tau_1(x, z)) \frac{m-2}{2} \left[\frac{1}{\rho^{2m-4}} \Gamma^{\frac{1}{2}(m-4)} - \Gamma^{-\frac{1}{2}m} \right] \\ + C'(x) \left[\Gamma^{\frac{1}{2}(2-m)} + \frac{1}{\rho^{2m-4}} \Gamma^{\frac{1}{2}(m-2)} - \frac{2}{\rho^{m-2}} \right] + (m-2) \left[\frac{m-4}{\rho^{2m-4}} \Gamma^{\frac{1}{2}(m-4)} + m \Gamma^{-\frac{1}{2}m} \right].$$

Thus

$$(4.7) \quad \Gamma^{\frac{1}{2}m} G(v) = -\left(\frac{m-2}{2}\right) \tau_1(x, z) + \epsilon_1(x, z) + C'(x)(1 + \epsilon_2(x, z))\Gamma(x, z),$$

where

$$(4.7a) \quad \epsilon_1(x, z) = (2(m-2) + \frac{1}{2}\tau_1(x, z)) \frac{m-2}{\rho^{2m-4}} \Gamma^{m-2}, \\ \epsilon_2(x, z) = \frac{1}{\rho^{2m-4}} \Gamma^{m-2} - \frac{2}{\rho^{m-2}} \Gamma^{\frac{1}{2}m-1} \quad (\tau_1(x, z) \text{ from (4.6)-(4.6a)}).$$

THEOREM 4.1. *The kernel of the integral equation (3.37) is of the form*

$$K(x, z) = \frac{1}{k(z)} G(v),$$

where $k(z)$ is given by (3.33a) (cf. (3.34)–(3.36)) and $G(v)$ is of the form (4.7), (4.7a); moreover, $C(x)$ enters in $C'(x)$ (cf. (4.3)) only.

In view of the preceding developments and in particular in consequence of (4.6b) – (4.6e) it is inferred that, when the $A_{i,j}$ are constants, while the $B_i \equiv 0$, we have

$$\tau_{1,j}(z) \equiv 0 \quad (j = 1, \dots, m)$$

and

$$\tau_1(x, z) = \sum_j \delta_j(z, x)(x_j - z_j) \quad (\text{cf. (4.6a)});$$

a further examination will show that, in this case, $\tau_1(x, z)$ is of the order of magnitude of $r^2 (= (x_1 - z_1)^2 + \dots + (x_m - z_m)^2)$.

We recall that by (3.15) and (3.18)

$$(4.8) \quad \frac{r}{\sqrt{\Gamma}} \leq \frac{1}{h_0(z)} \quad (0 \leq r \leq \delta_0).$$

Now, in consequence of (4.7) and (4.7a)

$$|\Gamma^{1/m} G(v)| \leq r k_1(z, \delta_0) \quad (r \leq \delta_0),$$

where $k_1(z, \delta_0)$ is independent of (x) . Hence from (4.8) it is deduced that

$$(4.9) \quad |G(v)| \leq k(z, \delta_0) r^{-m+1} \quad (r \leq \delta_0),$$

where

$$k(z, \delta_0) = \frac{k_1(z, \delta_0)}{h_0^m(z)}$$

is independent of (x) .

Choosing for the element of volume, dT_x , the expression of (3.7a), in view of (4.9) it is concluded that the integral

$$\int^{(m)} G^2(v) dT_x,$$

extended over a neighborhood of $(x) = (z)$, will not necessarily exist. That is, the integral

$$\int^{(m)} K^2(x, z) dT_x$$

may be divergent.

We shall consider now the effect of iterating the equation (3.37). It is inferred that the result of a ν -fold iteration is

$$(4.10) \quad u(z) - \int_D^{(m)} K_\nu(x, z) u(x) dT_x = f_{\nu,r}(z) \quad (\nu = 1, 2, \dots),$$

where

$$(4.11) \quad K_\nu(x, z) = \int^{(m)} K(x, x_1) K_{\nu-1}(x_1, z) dT_{x_1} \quad (\nu = 1, 2, \dots; K_0(x, z) = K(x, z))$$

and

$$(4.11a) \quad f_{\rho, \nu}(z) = f_{\rho, \nu-1}(z) + \int^{(m)} K_{\nu-1}(x_1, z) f_{\rho}(x_1) dT_{x_1} \\ (\nu = 1, 2, \dots; f_{\rho, 0}(z) = f_{\rho}(z)).$$

In view of a known theorem¹¹ it is concluded that, if $K(x, z)$ is of the form

$$(4.12) \quad K(x, z) = \frac{H(x, z)}{r^{\alpha}} \quad (r^2 = (x_1 - z_1)^2 + \dots + (x_m - z_m)^2; 0 < \alpha < m),$$

where $H(x, z)$ is bounded (integrable), then $K_{\nu}(x, z)$ is bounded (for (x) , (z) in Γ_{ρ}) for ν such that $\nu + 1$ is the least integer for which

$$\nu + 1 > \frac{1}{1 - \frac{\alpha}{m}};$$

that is, for ν equal to the least integer for which

$$\nu > \frac{\alpha}{m - \alpha}.$$

Now, in the actual case under consideration it is observed that, in view of (4.9), one has

$$(4.13) \quad |K(x, z)| \leq \frac{1}{k(z)} |G(v)| \leq \frac{1}{k(z)} k(z, \delta_0) r^{-m+1};$$

thus, (4.12) holds with $\alpha = m - 1$. Accordingly, the function $K_{\nu}(x, z)$ of (4.11) is bounded (and can be shown to be integrable in (x) over D) for $\nu = m$ and for (x) in D . The upper bound of $|K_{\nu}(x, z)|$ will in general depend on (z) .

The integrals defining the $f_{\rho, \nu}(z)$ (4.11a) will certainly exist since $f(x)$ is bounded in any closed subset of D and is integrable over Γ_{ρ} . This fact is a consequence of the following considerations. By (3.6) and since

$$\Gamma(x, z) \leq \rho^2 \quad (\text{for } (x) \text{ in } \Gamma_{\rho})$$

one has

$$v(x, z) = \Gamma^{1(2-m)}(x, z) v_0(x, z),$$

where

$$|v_0(x, z)| \leq 4 \quad ((x) \text{ in } \Gamma_{\rho}).$$

Furthermore

$$v(x, z) = w(x, z)^{m-2} v_0(x, z) r^{-m+2} \quad (\text{cf. (3.15)})$$

and, by virtue of (3.18),

$$(4.14) \quad |v(x, z)| \leq h'(z) r^{-m+2} \quad (h'(z) = 4h_0^{-m+2}(z))$$

¹¹ See, for instance, V. Volterra and J. Pérès, *Théorie générale des fonctionnelles* . . . [Paris, 1936; p. 275].

for (x) in Γ_ρ . On taking

$$dT_s = r^{m-1} dr d\Omega$$

and on using (4.14), we obtain for the integrand of (3.37a) the inequality

$$\left| \frac{v(x, z)}{k(z)} f(x) dT_s \right| \leq \frac{h'(z)}{k(z)} |f(x)| r dr d\Omega.$$

Whence the integral (3.37a) exists. In particular, if

$$(4.15) \quad |f(x)| \leq M \quad (\text{in } \Gamma_\rho)$$

one has

$$(4.16) \quad |f_\rho(z)| \leq \frac{h'(z)}{2k(z)} ML^2 S_m \leq f_0^*(z) \quad (L = \text{diameter of } \Gamma_\rho).$$

In view of (3.36), (4.14) we may take

$$(4.17) \quad f_0^*(z) = \frac{2ML^2}{m-2} H^{4m}(z) h_0^{-m+2}(z) \quad (\text{cf. (3.35)}).$$

Here $h_0^2(z) = h(z) - \delta_0 \lambda(2)$, where $h(z)$ is the minimum of $H(z, \pi)$ for (π) on the unit hypersphere; $\lambda(z)$ (from (3.16d)) may be taken as $\sigma \lambda_0(z)$ (suitable positive constant σ), with $\lambda_0(z)$ from (3.16c). In this connection, $\rho(>0)$ is to be taken sufficiently small so that, for (x) in Γ_ρ , one has $r \leq \delta_0 < h(z)/\lambda(z)$.

Let

$$[z]$$

denote generically a positive function of (z) uniformly bounded in every closed subset of D . On taking account of (4.11a) (with $\nu = 1$) and of the fact that, in consequence of (4.13)

$$|K(x_1, z)| \leq [z] r_1^{-m+1}$$

(r_1 = distance between (z) and (x_1)) we have (on taking $dT_{x_1} = r_1^{m-1} dr_1 d\Omega_1$)

$$|K(x_1, z) f_\rho(x_1) dT_{x_1}| \leq [z][x_1] dr_1 d\Omega_1,$$

it is concluded that $|f_\rho(z)| = [z]$. In view of (4.11a) (with $\nu = 2$) and since

$$|K_1(x_1, z) f_\rho(x_1) dT_{x_1}| \leq [z][x_1] r_1^{-m_1} dT_{x_1} \quad (-m_1 \geq -m + 1),$$

so that

$$|K_1(x_1, z) f_\rho(x_1) dT_{x_1}| \leq [z][x_1] r_1^{-m_1+m-1} dr_1 d\Omega_1,$$

we infer that, inasmuch as $-m_1 + m - 1 \geq 0$, necessarily $|f_{\rho,2}(z)| = [z]$. Thus, on using the fact that

$$|K_r(x, z)| \leq [z] r^{-m_r},$$

where

$$-m + 1 \leq -m_1 \leq -m_2 \leq \dots,$$

from (4.11a) we deduce, step by step, that

$$|f_{\rho, \nu}(z)| = [z]$$

for $\nu = 1, \dots, m$. In the above we use the fact that for every (z) in D , there is a closed subset of D (containing (z)) so that $K_\nu(x, z)$ is zero for (x) exterior this subset.

THEOREM 4.2. *Every solution u of the partial differential equation (1.1), which together with the $D_1 u$ is continuous in the domain D , satisfies the integral equation*

$$(4.18) \quad u(z) - \int_{\Gamma_{\rho'}}^{(m)} K_m(x, z) u(x) dT_x = f_{\rho, m}(z) \quad (\rho' > \rho),$$

where $K_m(x, z)$ and $f_{\rho, m}$ are defined with the aid of the relations (4.11), (4.11a). This equation has the advantage over the original equation (3.37) in the fact that the kernel $K_m(x, z)$ satisfies the inequality

$$|K_m(x, z)| \leq k^*(z) < \infty \quad (k^*(z) \text{ independent of } (x))$$

for (x) in $\Gamma_{\rho'}$, and that $K_m(x, z)$ is integrable, in (x) , over $\Gamma_{\rho'}$.

NOTE. By taking $\rho = \rho(z) (> 0)$ suitably small we arrange to have $\rho' = \rho'(z) > \rho(z)$, above, so small that the domain $\Gamma_{\rho'}$ (i.e. the set of points (x) such that $\Gamma(x, z) \leq (\rho')^2$) lies in D . We have $K_m(x, z)$ zero for (x) (in D) exterior $\Gamma_{\rho'}$; in (4.18) we may then replace the field of integration by D . The occurrence of ρ' takes place in consequence of the following considerations. By (4.11) (with $\nu = 1$)

$$(4.19) \quad K_1(x, z) = \int_D^{(m)} K(x, x_1) K(x_1, z) dT_{x_1} = \int_{\Gamma_\rho}^{(m)} K(x, x_1) K(x_1, z) dT_{x_1}$$

since $K(x_1, z) = 0$ for (x_1) exterior Γ_ρ . Now $K(x, x_1) = 0$ for (x) (in D) such that $\Gamma(x, x_1) \geq \rho(x_1) > 0$. There exists a number $\rho_1 = \rho_1(z) > \rho(z)$, independent of (x_1) ((x_1) in Γ_ρ) so that $K(x, x_1) = 0$ for (x) (in D) exterior Γ_{ρ_1} , for all (x_1) in Γ_ρ . Hence $K_1(x, z) = 0$ for (x) exterior Γ_{ρ_1} ; $\rho_1(z)$ can be made as small as desired by suitable choice of $\rho = \rho(z)$. By (4.11) (with $\nu = 2$)

$$(4.19a) \quad K_2(x, z) = \int_D^{(m)} K(x, x_1) K_1(x_1, z) dT_{x_1} = \int_{\Gamma_{\rho_1}}^{(m)} K(x, x_1) K_1(x_1, z) dT_{x_1}.$$

Repeating the above reasoning, with ρ_1 in place of ρ we find that $K_2(x, z) = 0$ for (x) exterior Γ_{ρ_2} , where $\rho_2 = \rho_2(z) (> \rho_1(z))$. Proceeding step by step in such a manner, we obtain the stated property of $K_m(x, z)$, with $\rho' = \rho(z)$.

In the sequel the prime over ρ will be deleted.

If the D_n are domains, in D , such that $\bar{D}_n \subset D_{n+1}$ and $D_n \rightarrow D$ (as $n \rightarrow +\infty$), with the aid of the above we obtain the following result:

$$(4.20) \quad u(z) - \int_{D_\infty}^{(m)} K_m(x, z) u(x) dT_x = f_{\rho, m}(z)$$

for all (z) in D_n ; here n' is some integer $> n$ such that $D_{n'}$ contains every point (x) for which $\Gamma(x, z) \leq \rho^2(z)$, when (z) is any point in \bar{D}_n ; with $\rho(z)$ suitably small such n' certainly exists and, in fact, may be taken as $n + 1$. Clearly $K_m(x, z) = 0$ for (z) in D_n and (x) in $D - \bar{D}_n$.

We have $k^*(z) \leq b_n < \infty$ (for (z) in \bar{D}_n), where $k^*(z)$ is from (4.18a). Thus

$$(4.20a) \quad |K_m(x, z)| \leq b_n < \infty \quad (\text{constant } b_n)$$

for all (x) in D and for all (z) in D_n .

5. Characteristic values and functions

In place of (1.1) we shall now consider the partial differential equation

$$(5.1) \quad F(u) + \lambda u = f \quad (F \text{ from (1.1); continuous } f \in L_2 \text{ in } D),$$

where λ is a parameter. This equation is obtained by replacing C by $C + \lambda$. Any equation of the form

$$F(u) + \lambda q(x)u = f$$

is reduced to the form (5.1), provided that, for (x) in D , the functions

$$A_{i,k}\sigma, \quad B_i\sigma, \quad C\sigma, \quad f\sigma \quad \left(\text{with } \sigma = \frac{1}{q(x)} \right)$$

satisfy conditions as previously imposed on the functions

$$A_{i,k}, B_i, C, f,$$

respectively.

The introduction of a parameter, as above, is in agreement with the situation in the Schrödinger wave theory.

On taking account of (3.37b), $v(x, z)$ being of the form (3.6), and of (2.1), replacing C by $C + \lambda$ we obtain

$$(5.2) \quad K(x, z) = K^0(x, z) + \frac{\lambda}{k(z)} v(x, z),$$

where $K^0(x, z)$ stands for the function which in section 4 has been designated as $K(x, z)$.

The function $K(x, z)$, defined in (5.2), is the kernel of the integral equation (3.37), corresponding to the partial differential equation (5.1).

Application of (4.11), for $v = 1, \dots, m$, will yield

$$(5.3) \quad K_m(x, z) = K_m^0(x, z) + \sum_{j=1}^{m+1} \lambda^j A_{m,j}(x, z).$$

Here $K_m^0(x, z)$ is the function designated in section 4 as $K_m(x, z)$. The $A_{m,j}(x, z)$ are functions of the same type as $K_m(x, z)$ of section 4 and accordingly satisfy inequalities of the form (4.20a).

In view of (3.37a), (4.11) and (4.11a) it is observed that $f_{p,m}(z)$ is a polynomial

in λ of degree m , whose coefficients are uniformly bounded (in absolute value) and integrable over Γ_p . In order to put in evidence dependence on λ we shall write

$$f_{p,m}(z) = f_{p,m}(z, \lambda).$$

The integral equation corresponding to (5.1) is

$$(5.4) \quad u(z) - \int_{\Gamma_p}^{(m)} u(x) \left[K_m^0(x, z) + \sum_{j=1}^{m+1} \lambda^j A_{m,j}(x, z) \right] dT_x = f_{p,m}(z, \lambda).$$

Before proceeding further we shall note that in consequence of a recent work of G. Giraud,¹² in the sequel referred to as (G), the following may be asserted with regard to the problem

$$(5.5) \quad F(u) + \lambda u = f \quad ((x) \text{ in } D_n),$$

$$(5.5a) \quad \frac{\partial u}{\partial \nu} + a_n(x)u = p_n(x) \quad ((x) \text{ on } S_n),$$

where D_n is a domain, for which $\bar{D} \subset D$, and S_n is the limiting surface of D_n ; the functions $a_n(x)$, $p_n(x)$ are defined and continuous on S_n . It is assumed that D_n and S_n are "regular" in the sense that S_n can be covered by a finite number of domains, in each of which one of the Cartesian coördinates of a point (variable in S_n) is expressible with the aid of $m - 1$ other coördinates through a function whose partial derivatives of the first order belong to the class Lip. h (with $h \leq 1$).¹³

In (G) the problem (5.5), (5.5a) (and, in fact, a more general problem) is reduced to integral equations to which the Fredholm theory applies. The final conclusion of (G), needed in the sequel, is that "either the only solution of the homogeneous problem is zero, and then the non homogeneous problem is compatible for all f (continuous in $D_n + S_n$) and $p_n(x)$; or the homogeneous problem has solutions not identically zero, deducible from p (p finite) linearly independent solutions, and then the non homogeneous problem is not compatible, unless certain p necessary and sufficient conditions are satisfied." The latter conditions are supplied by the Fredholm theory.

We shall take $p_n(x) \equiv 0$. Associated with (5.5), (5.5a) there is a set of characteristic values,

$$(5.6) \quad \lambda_{n,i} \quad (i = 1, 2, \dots),$$

the sequence $\lambda_{n,1}, \lambda_{n,2}, \dots$ possessing no finite limiting values. When $\lambda \neq \lambda_{n,i}$,

¹² G. Giraud, *Nouvelle méthode pour traiter certains problèmes relatifs aux équations du type elliptique* [Journal de Mathématiques, Tome Dix-huitième (1939), 111-143].

¹³ This is the type of surfaces considered in (G).

then the non homogeneous problem (5.5), (5.5a) has a solution, $u = u_n$, continuous in $D_n + S_n$. For $\lambda = \lambda_{n,i}$ the homogeneous problem

$$(5.7) \quad \begin{aligned} F(u) + \lambda_{n,i} u &= 0 & (\text{in } D_n), \\ \frac{\partial u}{\partial \nu} + a_n(x) u &= 0 & (\text{on } S_n) \end{aligned}$$

has at least one solution, $u = u_{n,i}$, distinct from zero, continuous in $D_n + S_n$.

We assume that D is such that there exists a sequence of domains

$$D_n \quad (n = 1, 2, \dots; \bar{D}_n \subset D),$$

whose limiting surfaces are S_n ($n = 1, 2, \dots$), respectively, and which are "regular" in the sense specified above, while

$$\bar{D}_n \subset D_{n+1}, \quad \lim_n D_n = D.^{14}$$

Forthwith the D_n and S_n will have the meaning just indicated.

In connection with (5.1) one may consider the "adjoint" non homogeneous equation

$$G(v) + \lambda v = g \quad (\text{continuous } g \subset L_2 \text{ in } D).$$

With respect to this equation results will hold precisely analogous to those stated from (5.1) to (5.4) for the equation (5.1).

Furthermore, applying the work of Giraud to the problem

$$(5.8) \quad \begin{aligned} G(v) + \rho v &= g & (\text{in } D_n), \\ \frac{\partial v}{\partial \nu} + b_n(x) v &= q_n(x) & (\text{on } S_n), \end{aligned}$$

where $b_n(x)$, $q_n(x)$ are continuous on S_n , we have a situation similar to that described in connection with (5.5), (5.5a). In particular, the problem (5.8) has a set of *characteristic values*

$$(5.9) \quad \rho_{n,j} \quad (j = 1, 2, \dots),$$

the sequence $\rho_{n,1}, \rho_{n,2}, \dots$ having no finite limiting points, such that the homogeneous problem

$$(5.10) \quad \begin{aligned} G(v) + \rho_{n,j} v &= 0 & (\text{in } D_n), \\ \frac{\partial v}{\partial \nu} + b_n(x) v &= 0 & (\text{on } S_n) \end{aligned}$$

has a solution, $v = v_{n,j}$, distinct from zero, continuous in $D_n + S_n$.

¹⁴ Here $\lim D_n = D_1 + D_2 + \dots$.

Under suitable general conditions on u and v , in view of the fundamental formula (2.7) we have

$$(5.11) \quad \int_{D_n}^{(m)} [vF(u) - uG(v)] dT = - \int_{S_n}^{(m-1)} \left[v \frac{\partial u}{\partial \nu} - u \frac{\partial v}{\partial \nu} + L_n uv \right] dS,$$

where (cf. 2.7a)) L_n is defined on S_n by

$$(5.12) \quad L_n = \sum_i \pi_i(n) B_i - \sum_{i,k} \pi_i(n) \frac{\partial A_{i,k}}{\partial x_k};$$

here the $\pi_i(n)$ are direction cosines of normal to the surface S_n . In view of the previously made suppositions with respect to the B_i , $A_{i,k}$ and the surface S_n , the function L_n is continuous on S_n .

It is observed that, if v satisfies the second relation (5.10) and u satisfies the last equation (5.7), one has

$$v \frac{\partial u}{\partial \nu} - u \frac{\partial v}{\partial \nu} + L_n uv = (b_n - a_n + L_n) uv \quad (\text{on } S_n).$$

Hence in order to make the above expression vanish on S_n we shall take

$$(5.13) \quad b_n = a_n - L_n.$$

THEOREM 5.1. 1°. *The set of characteristic values $\{\lambda_{n,i}\}$ [(5.6)] is identical with the set of characteristic values $\{\rho_{n,i}\}$ [(5.9)].¹⁵ We write $\lambda_{n,i} = \rho_{n,i}$.*

2°. *When F is self adjoint the characteristic values are all real and the characteristic functions may be taken real.*

3°. *There are on hand orthogonality relations:*

$$(5.14) \quad \int_{D_n}^{(m)} u_{n,i}(x) v_{n,j}(x) dT = 0 \quad (\text{for } \lambda_{n,i} \neq \lambda_{n,j});$$

$$(5.14a) \quad \int_{D_n}^{(m)} u_{n,i}(\bar{x}) u_{n,j}(x) dT = 0 \quad (\text{for } \lambda_{n,i} \neq \lambda_{n,j}, \text{ when } F \text{ is self adjoint}).$$

To prove (1°) consider a characteristic value $\lambda_{n,i}$ for (5.7); thus,

$$(5.15) \quad \begin{aligned} F(u_{n,i}) + \lambda_{n,i} u_{n,i} &= 0 && (\text{in } D_n), \\ \frac{\partial u_{n,i}}{\partial \nu} + a_n(x) u_{n,i} &= 0 && (\text{on } S_n), \end{aligned}$$

where $u_{n,i}$ is a characteristic function. Suppose, if possible, that $\lambda_{n,i}$ is not a characteristic value for (5.10) (under (5.13)).¹⁶ Then every non homogeneous problem

$$\begin{aligned} G(v) + \lambda_{n,i} v &= g && (\text{in } D_n), \\ \frac{\partial v}{\partial \nu} + b_n v &= 0 && (\text{on } S_n), \end{aligned}$$

¹⁵ Nothing is said about the multiplicities of these values.

¹⁶ We take $n_1 < n_2 < \dots$; $n_s \rightarrow \infty$ with s .

where $\lambda_{n,i}$ is fixed but g is any function continuous in $D_n + S_n$, will have a solution v . Now substitute this, supposedly existent, function v , as well as $u = u_{n,i}$, in (5.11). One obtains

$$0 = \int_{D_n}^{(m)} [vF(u_{n,i}) - u_{n,i}G(v)] dT = - \int_{D_n}^{(m)} u_{n,i}g dT,$$

which implies that $u_{n,i} = 0$ in D_n , contrary to the assertion made subsequent to (5.7). Hence $\lambda_{n,i}$ is a characteristic value for the "associated" homogeneous boundary value problem. The converse is proved by a similar method.

Substitute in (5.11) $u = u_{n,i}$, $v = v_{n,j}$. In view of the remark preceding (5.13) the integrand of the second member in (5.11) is zero on S_n ; accordingly

$$(5.16) \quad 0 = \int_{D_n}^{(m)} [v_{n,j}F(u_{n,i}) - u_{n,i}G(v_{n,j})] dT = (\lambda_{n,j} - \lambda_{n,i}) \int_{D_n}^{(m)} u_{n,i}v_{n,j} dT.$$

This implies (5.14).

In view of (4.2) and (4.3) the case when F is self adjoint, that is when $F = G$, is on hand if and only if

$$(5.17) \quad B_k(x) = \sum_i \frac{\partial A_{i,k}(x)}{\partial x_i} \quad (\text{in } D).$$

Thus in the self adjoint case it follows by (5.12) that

$$L_n = 0.$$

Under (5.13), the homogeneous boundary value problems (5.7) and (5.10) become identical. In consequence of these considerations and of (5.14) it is inferred that (5.14a) holds as stated.

With F self adjoint suppose, if possible, that there is a non real characteristic value $\lambda_{n,i}$. On taking conjugates of the members in (5.15) we deduce that $\bar{\lambda}_{n,i}$ will be another characteristic value, while $\bar{u}_{n,i}$ will be a characteristic function corresponding to $\bar{\lambda}_{n,i}$. By virtue of (5.16), where we put

$$\lambda_{n,j} = \bar{\lambda}_{n,i}, \quad v_{n,j} = \bar{u}_{n,i},$$

it is concluded that

$$(\bar{\lambda}_{n,i} - \lambda_{n,i}) \int_{D_n}^{(m)} |u_{n,i}|^2 dT = 0$$

and that, accordingly, $\bar{\lambda}_{n,i} - \lambda_{n,i} = 0$. This is impossible with $\lambda_{n,i}$ non real. Thus part 2° of the Theorem has been established.

In the self adjoint case the $u_{n,i}$ ($i = 1, 2, \dots$) will be always arranged as a set orthonormal in D_n ; we define

$$u_{n,i}(x) = 0 \quad (\text{in } D - (D_n + S_n));$$

the $u_{n,i}(x)$ will form a set orthonormal in D .

When F is not self adjoint orthonormalization of the $u_{n,i}$ ($i = 1, 2, \dots$) will yield a sequence whose members in general will not be characteristic func-

tion. This apparently involves the main reason for the difficulty of the treatment of equations which are not necessarily self adjoint—at least in certain questions relating to such equations.

The set of numbers

$$\lambda_{n,i} \quad (n, i = 1, 2, \dots)$$

is at most denumerably infinite. Hence there exists a real number λ_0 distinct from all of the above numbers. One may write

$$F(u) + \lambda u \equiv F^*(u) + \lambda^* u, \quad G(u) + \lambda u \equiv G^*(u) + \lambda^* u,$$

where

$$F^* = F + \lambda_0, \quad G^* = G + \lambda_0, \quad \lambda^* = \lambda - \lambda_0.$$

This amounts to augmenting the coefficients of u in F and G by λ_0 . In view of these considerations, augmenting, if necessary, C (and, hence $C' = C'' + C$, where C'' is independent of C) by a suitable λ_0 and retaining the original notation, we may consider that the $\lambda_{n,i}$ ($n, i = 1, 2, \dots$) are all distinct from zero; clearly, this entails no loss of generality.

In the sequel we confine ourselves to the case when F is self adjoint and when, accordingly, (5.17) holds.

6. The non homogeneous self adjoint problem

Let us first consider equation (5.1) for a non real value of λ . Then, in view of the reality of the characteristic values (cf. 2° of Theorem 5.1) of the problem (5.7), it can be asserted that the problem (with continuous f possibly complex valued, $\subset L_2$ in D)

$$(6.1) \quad \begin{aligned} F(u) + \lambda u &= f && (\text{in } D_n), \\ \frac{\partial u}{\partial \nu} + a_n(x)u &= 0 && (\text{on } S_n) \end{aligned}$$

has a continuous solution u_n in $D_n + S_n$. The function \bar{u}_n will satisfy

$$\begin{aligned} F(\bar{u}_n) + \bar{\lambda}\bar{u}_n &= \bar{f} && (\text{in } D_n), \\ \frac{\partial \bar{u}_n}{\partial \nu} + a_n(x)\bar{u}_n &= 0 && (\text{on } S_n). \end{aligned}$$

Accordingly application of (5.11) to

$$u = u_n, \quad v = \bar{u}_n$$

will yield

$$0 = \int_{D_n}^{(m)} [\bar{u}_n F(u_n) - u_n F(\bar{u}_n)] dT$$

and, hence,

$$\int_{D_n}^{(m)} |u_n|^2 dT = \frac{1}{(\lambda - \bar{\lambda})} \int_{D_n}^{(m)} (\bar{u}_n f - u_n \bar{f}) dT.$$

By the Schwarzian inequality

$$\int_{D_n}^{(m)} |u_n|^2 dT \leq \frac{2}{|\lambda - \bar{\lambda}|} \left[\int_{D_n}^{(m)} |u_n|^2 dT \right]^{\frac{1}{2}} \left[\int_{D_n}^{(m)} |f|^2 dT \right]^{\frac{1}{2}}$$

and, finally,

$$(6.2) \quad \int_{D_n}^{(m)} |u_n|^2 dT \leq \frac{4}{|\lambda - \bar{\lambda}|^2} \int_{D_n}^{(m)} |f|^2 dT \leq \frac{4}{|\lambda - \bar{\lambda}|^2} \int_D^{(m)} |f|^2 dT = \alpha(\lambda)$$

($n = 1, 2, \dots$; $\alpha(\lambda)$ independent of n and $< +\infty$). We define

$$(6.3) \quad u_n(x) = 0 \quad (\text{in } D - (D_n + S_n)).$$

Then

$$(6.4) \quad \int_D^{(m)} |u_n|^2 dT \leq \alpha(\lambda) \quad (n = 1, 2, \dots; \text{cf. (6.2)}).$$

This is analogous to a procedure in (C).

In consequence of a theorem of F. Riesz, inequalities of the form (6.4) imply that there exists an infinite subsequence

$$u_{n_i}(x) \quad (i = 1, 2, \dots)$$

and a function $u(x)$, for which

$$(6.5) \quad \int_D^{(m)} |u(x)|^2 dT \leq \alpha(\lambda)$$

such that

$$u_{n_i}(x) \rightarrow u(x) \quad (\text{in } D; \text{ as } n_i \rightarrow \infty)$$

in the *weak* sense.

Now, inasmuch as $u_{n_i}(x)$ is a solution of (5.1), in D_{n_i} , u_{n_i} satisfies the integral equation (5.4); thus with $r > i$

$$(6.6) \quad u_{n_r}(z) = \int_{D_{n_{i+1}}}^{(m)} u_{n_r}(x) \left[K_m^0(x, z) + \sum_{s=1}^{m+1} \lambda^s A_{m,s}(x, z) \right] dT_x + f_{\rho,m}(z, \lambda)$$

for (z) in D_{n_i} ($\rho(z) > 0$, being suitably chosen so that the kernel vanishes for (x) in $D - \bar{D}_{n_{i+1}}$, when (z) is in D_{n_i}).

We recall now a theorem of F. Riesz which, for a single variable, asserts that if

$$f_\nu(x) \in L_2 \quad (\text{on } (a, b); \nu = 1, 2, \dots), \quad f_\nu(x) \rightarrow f(x) \quad (\text{weakly})$$

and if

$$g(x) \in L_2 \quad (\text{on } (a, b))$$

then

$$\lim_{\nu} \int_a^b f_\nu(x) g(x) dx = \int_a^b f(x) g(x) dx.$$

An obvious extension of this theorem to the m -space of the point (x) enables us to assert that

$$(6.7) \quad \lim_r \int_{D_{n_i+1}}^{(m)} u_{n_r}(x) K_m(x, z) dT_x = \int_{D_{n_i+1}}^{(m)} u(x) K_m(x, z) dT_x$$

(for (z) in D_{n_i}), if

$$K_m(x, z) \subset L_2 \quad (\text{in } (x), \text{ over } D).$$

Now the latter property certainly holds by the preceding.

Thus, given any (z) in D , we choose i so that D_{n_i} contains (z) ; from (6.6) and (6.7) we then infer that

$$(6.8) \quad \lim u_{n_r}(z) = u'(z)$$

exists in the ordinary sense for (z) in D ; clearly

$$(6.9) \quad u'(z) = u(z) \quad (\text{in } D)$$

and

$$u(z) - \int_D^{(m)} u(x) K_m(x, z) dT_x = f_{\rho, m}(z, \lambda).$$

In (6.9) $u(z)$ is the function for which inequality (6.5) has been asserted.

We have the following theorem.

THEOREM 6.1. *Consider the partial differential equation (5.1), with $f(x)$ continuous and $\subset L_2$ in D ; let F be self adjoint.*

For every non real value of the parameter λ (5.1) possesses a solution $u(x)$ such that

$$\int_D^{(m)} |u(x)|^2 dT \leq \alpha(\lambda). \quad (\alpha(\lambda) \text{ from (6.2)}).$$

This theorem is analogous to a result established in (C) for the case when (1.3) holds.

We now wish to investigate the more complicated situation when λ is allowed to be real. In this there will be a certain analogy with an earlier work of Trjitzinsky.⁴

For λ distinct from the $\lambda_{n,i}$ ($i = 1, 2, \dots$) the problem (6.1) has a solution $u_n(x)$. On substituting $u = u_n$, $v = u_{n,i}$ in (5.11) and noting (5.15) and the fact that $L_n = 0$ we obtain

$$\int_{D_n}^{(m)} [u_{n,i} F(u_n) - u_n F(u_{n,i})] dT = 0$$

and

$$\int_{D_n}^{(m)} u_n(x) u_{n,i}(x) dT = \frac{1}{\lambda - \lambda_{n,i}} \int_{D_n}^{(m)} f(x) u_{n,i}(x) dT.$$

Thus, by (6.3)

$$(6.10) \quad \int_D^{(m)} u_n(x) u_{n,i}(x) dT = \frac{1}{\lambda - \lambda_{n,i}} \int_D^{(m)} f(x) u_{n,i}(x) dT.$$

It is noted that (6.10) constitutes a relation between the Fourier coefficients, with respect to the $u_{n,i}$ ($i = 1, 2, \dots$), of the functions $u_n(x)$ and $f(x)$.

Now an orthonormal set $\{u_{n,i}\}$ ($i = 1, 2, \dots$) is necessarily complete. This can be established by an extension of the familiar methods, involved in known analogous, though much simpler, boundary value problems.

On writing

$$(6.11) \quad u^{n,i} = \int_D^{(m)} u_n(y) u_{n,i}(y) dT,$$

we have

$$(6.11a) \quad \int_D^{(m)} |u_n(x)|^2 dT = \sum_i |u^{n,i}|^2.$$

In view of (6.10)

$$(6.12) \quad S_n^2 = \sum_i |u^{n,i}|^2 = \sum_i \left| \frac{f^{n,i}}{\lambda - \lambda_{n,i}} \right|^2, \quad f^{n,i} = \int_D^{(m)} f(x) u_{n,i}(x) dT.$$

DEFINITION 6.1. We designate by T a set of points λ , in the complex λ -plane, such that

$$(6.13) \quad \sum_i \left| \frac{f^{n,i}}{\lambda - \lambda_{n,i}} \right|^2 \leq B(\lambda) < +\infty \quad (n = n_1, n_2, \dots; f^{n,i} \text{ from (6.12)})^{16}$$

for all λ in T , the function $B(\lambda)$ being independent of n .

Let us form the complement, with respect to the λ -plane, of the set of points

$$(6.14) \quad \lambda_{n,i} \quad (n = n_1, n_2, \dots; i = 1, 2, \dots).$$

We define O as the set of interior points of this complement. Clearly O contains all the points not on the axis of reals. O will contain also some points on the axis of reals if the complement, with respect to the axis, of the set of points (6.14) contains interior (with respect to the axis) points—the latter points will belong to O .¹⁷ With δ positive, let O_δ be the subset of O consisting of points at the distance $\geq \delta$ from the frontier of O ; then O_δ will be a closed two dimensional set; moreover,

$$(6.15) \quad |\lambda - \lambda_{n,i}| \geq \delta \quad (n = n_1, n_2, \dots; i = 1, 2, \dots)$$

whenever λ is in O_δ . In view of (6.15) we have

$$\sum_i \left| \frac{f^{n,i}}{\lambda - \lambda_{n,i}} \right|^2 \leq \frac{1}{\delta^2} \sum_i |f^{n,i}|^2 \quad (\text{in } O_\delta)$$

and, in consequence of the second relation (6.12) as well as of Bessel's inequality,

$$\sum_i \left| \frac{f^{n,i}}{\lambda - \lambda_{n,i}} \right|^2 \leq \frac{1}{\delta^2} \int_D^{(m)} |f(x)|^2 dT \quad (\text{in } O_\delta).$$

¹⁷ It is conceivable to have the set (6.14) everywhere dense on the axis of reals, in which case, by definition, O will consist of the λ -plane minus the axis of reals. However, the set T of Definition 6.1 may have points on the axis of reals, even under these circumstances, provided the functions f are suitably chosen.

The set O , defined in connection with (6.14), is a set T of Definition 6.1 with

$$B(\lambda) = \frac{1}{\delta^2} \int_D^{(m)} |f(x)|^2 dT$$

for λ in O ; here δ is the distance from the point representing λ to the frontier of O .

An important feature of sets T of the special form O lies in the fact that they are independent of the choice of f —in so far as $f \in L_2$ (in D); this is not so for the general sets T . In any case, of course, T includes all the points not on the axis of reals.

By (6.12) and (6.13)

$$S_n^2 \leq B(\lambda) \quad (n = n_1, n_2, \dots; \text{ in } T)$$

and, in view of (6.11a),

$$(6.16) \quad \int_D^{(m)} |u_n(x)|^2 dT \leq B(\lambda) < +\infty \quad (n = n_1, n_2, \dots)$$

for λ in T . The second member, here, being independent of n , one may assert that the sequence (n_1, n_2, \dots) contains a subsequence (k_1, k_2, \dots) so that

$$u_{k_j}(x) \rightarrow u(x) \quad (\text{as } k_j \rightarrow \infty; \text{ in } D),$$

convergence being in the weak sense to a function $u(x)$ for which

$$\int_D^{(m)} |u(x)|^2 dT \leq B(\lambda).$$

By a reasoning employed before and on making use of the integral equation it is concluded that $u(x)$ is a solution of the non homogeneous problem (5.1) for λ in T .

We have the following Existence Theorem.

THEOREM 6.2. *Let F be self adjoint (in D). The $\lambda_{n,i}$ ($n = n_1, n_2, \dots$; $i = 1, 2, \dots$) are the characteristic numbers of a sequence of approximating boundary value problems; together with the characteristic functions $u_{n,i}$ the $\lambda_{n,i}$ satisfy (5.15). In terms of the $u_{n,i}(x)$ (u_{n1}, u_{n2}, \dots orthonormalized in D) we define numbers $f^{n,i}$ by the second relations (6.12). Let T be a set for which (6.13) of Definition 6.1 holds.*

Then the non homogeneous problem

$$F(u) + \lambda u = f$$

will have a solution u , in $D, \subset L_2$ in D for every λ in T .

DEFINITION 6.2. *It will be said that F is of class I on a set O , if every solution, $\subset L_2$ in D , of the homogeneous problem*

$$F(u) + \lambda u = 0$$

is zero in D —this being so for every λ in O . In the contrary case F will be said to be of class II.

Designation of classes I, II, here, is suggested by an analogous usage in Carleman's theory of integral equations.

It is to be recalled that corresponding to the equation (5.1) we have an iterated integral equation (5.4); the latter may be written in the form

$$(6.17) \quad u(z) - \int_D^{(m)} u(x) \sum_{i=0}^{m+1} \lambda^i A_{m,i}(x, z) dT_x = f_{p,m}(z, x) \quad (A_{m,0} = K_m^0).$$

On taking account of the text in connection with (4.20) we note that in (6.17) the field of integration may be taken as D_{n+1} if (z) is in \bar{D}_n . The expression

$$(6.18) \quad a(z | D) = \sum_{i=0}^{m+1} \left[\int_D^{(m)} A_{m,i}^2(z_1, z) dT_s \right]^{\frac{1}{2}}$$

defines a function for all (z) in D . In fact, let (z) be a fixed point in D . Then for some n the domain D_n contains (z) ; for all (z_1) in $D - \bar{D}_{n+1}$ we have $A_{m,i}(z_1, z) = 0$. Hence

$$\int_D^{(m)} A_{m,i}^2(z_1, z) dT_{s_1} = \int_{D_{n+1}}^{(m)} A_{m,i}^2(z_1, z) dT_{s_1}.$$

Here

$$|A_{m,i}(z_1, z)| \leq \gamma_n < \infty \quad (\text{all } (z) \text{ in } D_n; \text{ all } (z_1) \text{ in } D)$$

in view of the statement with respect to (4.20a). Thus the integral above exists.

7. Spectral theory

The "spectrum" of F , with respect to the boundary value problem

$$F(u) + \lambda u = 0 \quad (\text{in } D_n; \text{ parameter } \lambda),$$

$$(7.1) \quad \frac{\partial u}{\partial \nu} + a_n(x)u = 0 \quad (\text{on } S_n),$$

we define to be the function $\theta_n(x, y | \lambda)$ for which

$$(7.2) \quad \begin{aligned} \theta_n(x, y | \lambda) &= \sum_{0 < \lambda_{n,i} < \lambda} u_{n,i}(x) u_{n,i}(y) & (\text{for } \lambda > 0), \\ \theta_n(x, y | \lambda) &= - \sum_{\lambda \leq \lambda_{n,i} < 0} u_{n,i}(x) u_{n,i}(y) & (\text{for } \lambda < 0), \end{aligned}$$

while $\theta_n(x, y | 0) = 0$; this function is defined for $(x), (y)$ in $D_n + S_n$ and for all real λ . Clearly $\theta_n(x, y | \lambda) = \theta_n(y, x | \lambda)$. Such definition is possible inasmuch as it had been previously arranged to have the $\lambda_{n,i}$ all distinct from zero and since the $\lambda_{n,i}$ are all real.

In consequence of the definition of $u_{n,i}$ we have

$$\theta_n(x, y | \lambda) = 0$$

for (x) exterior $D_n + S_n$, as well as for (y) exterior $D_n + S_n$.

From the definition of θ_n it follows that

$$\Delta_{\lambda_1}^{\lambda_2} \theta_n \equiv \theta_n(x, y | \lambda_2) - \theta_n(x, y | \lambda_1) = \sum_{\lambda_1 \leq \lambda_{n,i} < \lambda_2} u_{n,i}(x) u_{n,i}(y)$$

when $\lambda_2 > \lambda_1$. Thus, on taking

$$\lambda_0 < \lambda_1 < \dots < \lambda_q,$$

we obtain

$$(7.3) \quad \delta_n = \sum_{j=1}^q |\Delta_{\lambda_{n-1}}^{\lambda_j} \theta_n(x, y | \lambda)| \leq \sum_{\lambda_0 \leq \lambda_n, i < \lambda_q} |u_{n,i}(x) u_{n,i}(y)|$$

for (x) and (y) in D .

In view of the first relation (5.15) and of the previously established connection with integral equations we have

$$(7.4) \quad u_{n,i}(z) = \sum_{j=0}^{m+1} \lambda_{n,i}^j \int_{D_n}^{(m)} u_{n,i}(x) A_{m,i}(x, z) dT_x \quad [A_{m,0} = K_m^0]$$

for (z) in \bar{D}_{n-1} (if $\rho = \rho(z)$, > 0 , is suitably chosen). The functions in the kernel of (7.4) have the general properties indicated subsequent to (5.3).

In consequence of (7.3) and (7.4) one has

$$(7.5) \quad \delta_n \leq \sum_{\lambda_0 \leq \lambda_n, i < \lambda_q} \left| \sum_{j_1=0}^{m+1} \lambda_{n,i}^{j_1} \int_{D_n}^{(m)} u_{n,i}(z_1) A_{m,i_1}(z_1, x) dT_{z_1} \right| \cdot \left| \sum_{j_2=0}^{m+1} \lambda_{n,i}^{j_2} \int_{D_n}^{(m)} u_{n,i}(z_2) A_{m,i_2}(z_2, y) dT_{z_2} \right|$$

for (x) and (y) in \bar{D}_{n-1} . Here (as well as in (7.4)) D_n may be replaced by D . On letting λ^* denote the greater of the numbers 1, $|\lambda_0|$, $|\lambda_q|$ from (7.5) it is inferred that

$$\delta_n \leq (\lambda^*)^{2m+2} \sum_{j_1, j_2} \left\{ \sum_i \left| \int_D^{(m)} A_{m,i_1}(z_1, x) u_{n,i}(z_1) dT_{z_1} \right| \cdot \left| \int_D^{(m)} A_{m,i_2}(z_2, y) u_{n,i}(z_2) dT_{z_2} \right| \right\} \quad ((x), (y) \text{ in } \bar{D}_{n-1}).$$

Thus

$$\delta_n \leq (\lambda^*)^{2m+2} \sum_{j_1, j_2} \left\{ \left[\sum_i \left| \int_D^{(m)} A_{m,i_1}(z_1, x) u_{n,i}(z_1) dT_{z_1} \right|^2 \right]^{\frac{1}{2}} \cdot \left[\sum_i \left| \int_D^{(m)} A_{m,i_2}(z_2, y) u_{n,i}(z_2) dT_{z_2} \right|^2 \right]^{\frac{1}{2}} \right\}$$

for $(x), (y)$ in \bar{D}_{n-1} . On making use of Bessel's inequality we obtain

$$(7.7) \quad \delta_n \leq (\lambda^*)^{2m+2} \sum_{j_1, j_2} \left[\int_D^{(m)} A_{m,i_1}^2(z_1, x) dT_{z_1} \right]^{\frac{1}{2}} \cdot \left[\int_D^{(m)} A_{m,i_2}^2(z_2, y) dT_{z_2} \right]^{\frac{1}{2}} = (\lambda^*)^{2m+2} a(x, y)$$

$((x), (y) \text{ in } \bar{D}_{n-1})$, with

$$(7.7a) \quad a(x, y) = a(x | D) a(y | D),$$

where $a(x | D)$ is defined by (6.18).

It is essential to note that the last member in (7.7) is independent of the mode of subdivision of the interval (λ_0, λ_q) and is independent of n .

By (7.7) and in view of the definition of δ_n , given in (7.3), one has

$$(7.8) \quad V_\alpha^\beta \theta_n(x, y | \lambda) \leq (\lambda^*)^{2m+2} a(x, y)$$

for $(x), (y)$ in D_{n-1} . Here V_α^β denotes variation, with respect to λ , on the interval (α, β) ; λ^* is the greatest of the numbers $|\alpha|, |\beta|, 1$. Also, since $\theta_n(x, y | 0) = 0$ we have

$$|\theta_n(x, y | \lambda)| = |\Delta_0^\lambda \theta_n|$$

so that, in view of (7.7),

$$(7.9) \quad |\theta_n(x, y | \lambda)| \leq (\lambda')^{2m+2} a(x, y)$$

$((x), (y))$ in \bar{D}_{n-1} where λ' is the greater of the numbers 1, $|\lambda|$.

For any integer $r \geq n$, corresponding to (7.4), we have

$$(7.10) \quad u_{r,i}(z) = \sum_{j=0}^{m+1} \lambda_{n,i}^j \int_{D_n}^{(m)} u_{r,i}(x) A_{m,i}(x, z) dT_x$$

for (z) in \bar{D}_{n-1} . This is a consequence of the fact that D_r ($r \geq n$) contains D_n and that for (x) in $D - \bar{D}_n$

$$(7.11) \quad A_{m,i}(x, z) = 0 \quad (\text{for } (z) \text{ in } \bar{D}_{n-1}).$$

By a reasoning of the type employed subsequent to (7.4) for the purpose of derivation of (7.8) and (7.9) we now proceed from (7.10), (7.11) obtaining the inequalities

$$(7.12) \quad V_\alpha^\beta \theta_r(x, y | \lambda) \leq (\lambda^*)^{2m+2} a(x, y),$$

$$(7.13) \quad |\theta_r(x, y | \lambda)| \leq (\lambda')^{2m+2} a(x, y)$$

for $(x), (y)$ in \bar{D}_{n-1} and for $r = n + 1, \dots$

If the θ_r ($r \geq n$) had a property in the nature of equicontinuity in $(x), (y)$, for $(x), (y)$ in \bar{D}_{n-1} , then the "Compactness" theorem of Carleman² would be applicable. Under the existing conditions, however, a modification of the developments in (C₁; pp. 21-24), together with (7.12), (7.13), enables assertion of the following.

Given $n(> 0)$, there exists a subsequence

$$(7.14) \quad \theta_{r(n,1)}, \theta_{r(n,2)}, \dots$$

such that the limit

$$\lim_i \theta_{r(n,i)} = \theta(x, y | \lambda)$$

exists in \bar{D}_{n-1} , for real λ ; moreover,

$$(7.15) \quad V_\alpha^\beta \theta(x, y | \lambda) \leq (\rho^*)^{2m+2} a(x, y),$$

$$(7.15a) \quad |\theta(x, y | \lambda)| \leq (\rho')^{2m+2} a(x, y)$$

for (x) and (y) in \bar{D}_{n-1} , for all real λ , except perhaps for a denumerable infinity of values λ . Offhand, θ may depend on n ; whenever (x) or (y) is in $D - \bar{D}_{n-1}$, we define this function θ as zero.

We select a subsequence (7.14) for $n = 2$, obtaining a limiting function, ${}_1\theta(x, y | r)$ for $(x), (y)$ in \bar{D}_1 . From this subsequence we choose another subsequence

$$(7.16) \quad \theta_{r_{3,1}}, \theta_{r_{3,2}}, \theta_{r_{3,3}}, \dots$$

converging, for $(x), (y)$ in \bar{D}_2 , to a limiting function ${}_2\theta$, which is necessarily identical with ${}_1\theta$ for $(x), (y)$ in \bar{D}_1 .

We replace the sequence (7.16) by the sequence

$$(7.17) \quad \theta_{r(2,1)}, \theta_{r_{3,2}}, \theta_{r_{3,3}}, \theta_{r_{3,4}}, \dots$$

which has the same limit as the sequence (7.16). From (7.17) we select a sequence

$$\theta_{r(2,1)}, \theta_{r_{3,2}}, \theta_{r_{4,3}}, \theta_{r_{4,4}}, \theta_{r_{4,5}}, \dots$$

converging to ${}_3\theta$ in \bar{D}_3 ; clearly

$${}_3\theta = \begin{cases} {}_2\theta & \text{(for } (x), (y) \text{ in } \bar{D}_2), \\ {}_1\theta & \text{(for } (x), (y) \text{ in } \bar{D}_1). \end{cases}$$

Continuing this process of consecutive selections one obtains a sequence

$$\theta_{r(2,1)}, \theta_{r_{3,2}}, \theta_{r_{4,3}}, \theta_{r_{4,4}}, \dots$$

converging to a limiting function

$$\theta(x, y | \lambda)$$

for $(x), (y)$ in D , for all real λ , except perhaps for a denumerable infinity of values λ ; θ is independent of n and satisfies the conditions stated in connection with (7.15), (7.15a), this being so for all $(x), (y)$ in D .

DEFINITION 7.1. A function $\theta(x, y | \lambda)$ obtained as described above will be termed a spectrum of F , associated with the sequence of boundary value problems (7.1).

THEOREM 7.1. Associated with the sequence of boundary value problems (7.1) there exists at least one spectrum θ , satisfying the conditions stated in connection with (7.15), (7.15a) (cf. (7.7a)); we also note the statement preceding Definition 7.1.

In the remainder of this section we shall state a number of properties of $\theta(x, y | \lambda)$. These properties are analogous to those found, for functions designated as θ , by Carleman in Chapter I of (C₁). These results may be proved with the aid of the theorems referred to in (C₁; 7-24), as well as of Theorem 7.1. We shall omit the details of proof.

According to the preceding, there exists a sequence

$$\theta_{m_1}, \theta_{m_2}, \dots \quad (m_1 < m_2 < \dots)$$

such that

$$(7.19) \quad \lim \theta_{m_r} = \theta \quad ((x), (y) \text{ in } D; \text{ as } m_r \rightarrow \infty).$$

With (7.19) in view we shall write $n = m_r$. The following is asserted.

For almost all (x) in D

$$(7.20) \quad \begin{aligned} \psi(x, \lambda) &= \int_D^{(m)} \theta(x, y | \lambda) h(y) dT_y = \lim \psi_n(x, \lambda), \\ \psi_n(x, \lambda) &= \int_D^{(m)} \theta_n(x, y | \lambda) h(y) dT_y, \end{aligned}$$

whenever $h(y) \in L_2$ in D . If, in addition, $g \in L_2$ in D , then

$$\begin{aligned} \int_D^{(m)} g(x) \psi(x, \lambda) dT_x &= \lim \int_D^{(m)} g(x) \psi_n(x, \lambda) dT_x \\ &= \lim \int_D^{(m)} \int_D^{(m)} \theta_n(x, y | \lambda) g(x) h(y) dT_x dT_y; \end{aligned}$$

the order of integration in

$$(7.21) \quad \int_D^{(m)} \int_D^{(m)} \theta(x, y | \lambda) g(x) h(y) dT_x dT_y \quad (g, h \in L_2 \text{ in } D)$$

is immaterial.

On using (7.17) (for θ_n) it is inferred that

$$V_\alpha^\beta \psi_n(x, \lambda), \quad V_\alpha^\beta \psi(x, \lambda) \leq a_0 \left[\int_D^{(m)} |h(y)|^2 dT_y \right]^{\frac{1}{2}} a(x | D),$$

where $a_0 = a_0(\alpha, \beta)$ is finite for α, β finite. Moreover,

$$(7.22) \quad |\psi_n(x, \lambda)|, \quad |\psi(x, \lambda)| \leq a_1 a^2 \left[\int_D^{(m)} |h(y)|^2 dT_y \right]^{\frac{1}{2}},$$

where

$$(7.23) \quad a_1 = (\lambda')^{2m+2} \quad [\lambda' = \max. (1, |\lambda|)].$$

We have, whenever $g, h \in L_2$ in D ,

$$(7.24) \quad V_\alpha^\beta \int_D^{(m)} \int_D^{(m)} \theta_n(x, y | \lambda) g(x) h(y) dT_x dT_y$$

and

$$\begin{aligned} \left| \int_D^{(m)} \int_D^{(m)} \theta_n(x, y | \lambda) g(x) h(y) dT_x dT_y \right| \\ \leq \left[\int_D^{(m)} |g(x)|^2 dT_x \right]^{\frac{1}{2}} \left[\int_D^{(m)} |h(y)|^2 dT_y \right]^{\frac{1}{2}}; \end{aligned}$$

these inequalities will hold for $\theta(x, y | \lambda)$ as well.

When $\omega(\lambda)$ is continuous for $\lambda_0 \leq \lambda \leq \lambda_1$ one has

$$\int_{\lambda_0}^{\lambda_1} \omega(\lambda) d_\lambda \theta(x, y | \lambda) = \lim \int_{\lambda_0}^{\lambda_1} \omega(\lambda) d_\lambda \theta_n(x, y | \lambda),$$

$$\begin{aligned}
 \int_{\lambda_0}^{\lambda_1} \omega(\lambda) d\lambda \int_D^{(m)} h(y) \theta(x, y | \lambda) dT_y &= \lim \int_{\lambda_0}^{\lambda_1} \omega(\lambda) d\lambda \int_D^{(m)} h(y) \theta_n(x, y | \lambda) dT_y, \\
 \int_{\lambda_0}^{\lambda_1} \omega(\lambda) d\lambda \int_D^{(m)} \int_D^{(m)} \theta(x, y | \lambda) h(x) g(y) dT_x dT_y \\
 (7.25) \qquad \qquad \qquad &= \lim \int_{\lambda_0}^{\lambda_1} \omega(\lambda) d\lambda \int_D^{(m)} \int_D^{(m)} \theta_n(x, y | \lambda) h(x) g(y) dT_x dT_y
 \end{aligned}$$

for $(x), (y)$ in D , whenever $h, g \in L_2$ in D .

Also

$$\begin{aligned}
 \int_D^{(m)} (hy) \left[\int_{\lambda_0}^{\lambda_1} \omega(\lambda) d\lambda \theta(x, y | \lambda) \right] dT_y &= \int_{\lambda_0}^{\lambda_1} \omega(\lambda) d\lambda \left[\int_D^{(m)} \theta(x, y | \lambda) h(y) dT_y \right], \\
 \int_D^{(m)} \int_D^{(m)} g(x) h(y) \left[\int_{\lambda_0}^{\lambda_1} \omega(\lambda) d\lambda \theta(x, y | \lambda) \right] dT_x dT_y \\
 &= \int_{\lambda_0}^{\lambda_1} \omega(\lambda) \left[d\lambda \int_D^{(m)} \int_D^{(m)} \theta(x, y | \lambda) g(x) h(y) dT_x dT_y \right], \\
 \int_D^{(m)} g(x) \left\{ \int_{\lambda_0}^{\lambda_1} \omega(\lambda) d\lambda \left[\int_D^{(m)} \theta(x, y | \lambda) h(y) dT_y \right] \right\} dT_x \\
 &= \int_{\lambda_0}^{\lambda_1} \omega(\lambda) d\lambda \left[\int_D^{(m)} \int_D^{(m)} \theta(x, y | \lambda) g(x) h(y) dT_x dT_y \right].
 \end{aligned}$$

The integral

$$(7.26) \qquad \int_{\lambda_0}^{\lambda_1} d\lambda \int_D^{(m)} \int_D^{(m)} \theta(x, y | \lambda) \bar{h}(x) h(y) dT_x dT_y$$

converges for all $h \in L_2$ (in D). There exists a function $\psi(x)$, $\in L_2$ (in D), such that

$$\psi_l(x) = \int_{\lambda_0}^{\lambda_1} d\lambda \left[\int_D^{(m)} \theta(x, y | \lambda) h(y) dT_y \right] \rightarrow \psi(x)$$

(as $l \rightarrow +\infty$), convergence being in the mean square for (x) in D .

8. Developments on the basis of spectra

For a fixed $n = m$, consider the set of functions $u_{n,i}$ ($i = 1, 2, \dots$). Recalling the character of this sequence it is observed that the Parseval's equality for the $u_{n,i}$, expressed with the aid of a Stieltjes integral, is of the form

$$B_n = \int_{\lambda_0}^{\lambda_1} d\lambda \int_D^{(m)} \int_D^{(m)} \theta_n(x, y | \lambda) \bar{h}(x) h(y) dT_x dT_y \leq \int_D^{(m)} |h(y)|^2 dT_y.$$

Here h is any function, $\in L_2$ in D , and θ_n is the function introduced in (7.2). With $0 < l < \infty$, we have

$$(8.1) \quad B_{n,l} = \int_{\lambda_0}^{\lambda_1} d\lambda \int_D^{(m)} \int_D^{(m)} \theta_n(x, y | \lambda) \bar{h}(x) h(y) dT_x dT_y \leq \int_D^{(m)} |h(y)|^2 dT_y.$$

This may be asserted in view of the fact that

$$\int_{\alpha}^{\beta} d\lambda \int_D^{(m)} \int_D^{(m)} \theta_n(x, y | \lambda) \bar{h}(x) h(y) dT_x dT_y \geq 0$$

for all real α, β . In consequence of (7.25)

$$\lim_n B_{n,l} = \int_{-l}^l d\lambda \int_D^{(m)} \int_D^{(m)} \theta(x, y | \lambda) \bar{h}(x) h(y) dT_x dT_y.$$

Hence by (8.1)

$$\int_{-l}^l d\lambda \int_D^{(m)} \int_D^{(m)} \theta(x, y | \lambda) \bar{h}(x) h(y) dT_x dT_y \leq \int_D^{(m)} |h(y)|^2 dT_y.$$

Thus, on letting $l \rightarrow \infty$, we obtain (note convergence of the integral (7.26))

$$(8.2) \quad \int_{-\infty}^{\infty} d\lambda \int_D^{(m)} \int_D^{(m)} \theta(x, y | \lambda) \bar{h}(x) h(y) dT_x dT_y \leq \int_D^{(m)} |h(y)|^2 dT_y,$$

which is the *generalized Bessel's inequality* related to our problem; this inequality is associated with a certain sequence of homogeneous boundary value problems.

It will be convenient to introduce the following Definition.

DEFINITION 8.1. Suppose F is self adjoint. It will be said that F is closed with respect to a spectrum θ if

$$(8.3) \quad \int_{-\infty}^{\infty} d\lambda \int_D^{(m)} \int_D^{(m)} \theta(x, y | \lambda) \bar{h}(x) h(y) dT_x dT_y = \int_D^{(m)} |h(y)|^2 dT_y$$

for every $h \in L_2$ (in D).

It is observed that (8.3) is a particular instance of (8.2) and constitutes a *generalized Parseval's relation*.

Suppose now F is closed with respect to a spectrum θ . Let h be real valued. Replacing $h(x)$ in (8.3) by $a(x) + b(x)$, where $a(x), b(x) \in L_2$ (in D), in consequence of the closed character of F we obtain

$$(8.4) \quad \begin{aligned} & \int_{-\infty}^{\infty} d\lambda \int_D^{(m)} \int_D^{(m)} \theta(x, y | \lambda) a(x) b(y) dT_x dT_y \\ & + \int_{-\infty}^{\infty} d\lambda \int_D^{(m)} \int_D^{(m)} \theta(x, y | \lambda) b(x) a(y) dT_x dT_y = 2 \int_D^{(m)} a(y) b(y) dT_y. \end{aligned}$$

Interchanging (x) and (y) in the second term of the first member above, making use of the symmetry of θ and taking note of the permissibility of the interchange of order of integration in the integral (7.21), it is concluded that this term is equal to the first integral displayed in (8.4). Accordingly

$$(8.5) \quad \int_{-\infty}^{\infty} d\lambda \int_D^{(m)} \int_D^{(m)} \theta(x, y | \lambda) a(x) b(y) dT_x dT_y = \int_D^{(m)} a(y) b(y) dT_y.$$

Let C be a "cube" in the m -space, defined by inequalities

$$(8.6) \quad c_j \leq x_j \leq c_j + L \quad (j = 1, \dots, m; L > 0).$$

We choose the point $(c) = (c_1, \dots, c_m)$ and L sufficiently great so that

$$(8.6a) \quad D \subset C,$$

supposing that D is bounded. We define θ , for (x) and (y) in C , by the relations

$$(8.7) \quad \begin{aligned} \theta(x, y | \lambda) &= 0 && ((x) \text{ in } C - D; (y) \text{ in } C), \\ \theta(x, y | \lambda) &= 0 && ((y) \text{ in } C - D; (x) \text{ in } C). \end{aligned}$$

One then may express (8.5) in the form

$$(8.8) \quad \int_{-\infty}^{\infty} d\lambda \int_C^{(m)} \int_C^{(m)} \theta(x, y | \lambda) a(x) b(y) dT_x dT_y = \int_C^{(m)} a(y) b(y) dT_y,$$

where $a(x)$, $b(y)$ are any functions $\subset L_2$ in C , while

$$(8.8a) \quad a(x) = 0 \quad (\text{in } C - D).$$

Substitute, in particular,

$$b(y) = b_t(y) = b(t_1, \dots, t_m; y_1, \dots, y_m),$$

where the function $b_t(y)$ is defined, for (t) and (y) in C , by the relations

$$(8.9) \quad \begin{aligned} b_t(y) &= 1 && (c_j \leq y_j \leq t_j; j = 1, \dots, m), \\ b_t(y) &= 0 && (\text{elsewhere}).^{18} \end{aligned}$$

From (8.8) we then obtain

$$(8.10) \quad \int_{-\infty}^{\infty} d\lambda \sigma(t; \lambda) = \int_{y_1=c_1}^{t_1} \dots \int_{y_m=c_m}^{t_m} a(y) dT_y$$

where, in view of (8.8a),

$$(8.10a) \quad \sigma(t; \lambda) = \int_D^{(m)} a(x) \left\{ \int_{y_1=c_1}^{t_1} \dots \int_{y_m=c_m}^{t_m} \theta(x, y | \lambda) dT_y \right\} dT_x.$$

With the aid of (8.10) the following result is established.

THEOREM 8.1. *Suppose θ is a spectrum, with respect to which F is closed, and that D is bounded. There is then on hand the following generalized Fourier expansion*

$$(8.11) \quad a(t) = \frac{\partial^m}{\partial t_1 \dots \partial t_m} \int_{-\infty}^{\infty} d\lambda(t; \lambda)$$

for almost all (t) in D ; this is valid for all $a \subset L_2$ in D ; here $\delta(t; \lambda)$ is defined by (8.10a), with θ subject to (8.7) (defined in (8.6), (8.6a)).

An important application of spectra is, as will be now shown, in establishing explicit formulas for solutions of the non homogeneous problem (5.1).

We have

$$(8.12) \quad u_n(x) = \sum_i u^{n,i} u_{n,i}(x).$$

¹⁸ That is, for points (y) such that at least some y_j exceeds t_j ($(y), (t)$ in C).

Ordinary convergence, here, instead of convergence in the mean square takes place by reasoning of the same type as is used to prove the analogous fact in certain known simpler classical cases. As has been shown previously $u_n(x) \rightarrow u(x)$ (as $n = k_j \rightarrow \infty$) in the weak sense as well as in the ordinary sense. Certainly

$$(8.13) \quad \int_{(C)}^{(x)} u_{k_j}(x) dT_x \rightarrow \int_{(C)}^{(x)} u(x) dT_x$$

as $k_j \rightarrow \infty$, for (x) in C .¹⁹ Here and in the sequel

$$(8.14) \quad \int_{(C)}^{(x)} a(x) dT_x = \int_{t_1=C_1}^{x_1} \cdots \int_{t_m=C_m}^{x_m} a(t_1, \dots, t_m) dt_1 \cdots dt_m$$

((x) in C ; $a(x)$ integrable over C). Moreover, the functions $u_n(x)$ are taken equal to zero in $C - D$.

In view of (6.10) and of (6.11), from (8.12) it is deduced that

$$(8.15) \quad \begin{aligned} u_n(x) &= \sum_i \frac{1}{\lambda - \lambda_{n,i}} f^{n,i} u_{n,i}(x) = \sum_i \frac{1}{\lambda - \lambda_{n,i}} u_{n,i}(x) \int_D^{(m)} f(y) u_{n,i}(y) dT_y \\ &= \int_{-\infty}^{\infty} \frac{1}{\lambda - \rho} d_\rho \int_D^{(m)} f(y) \theta_n(x, y | \rho) dT_y. \end{aligned}$$

We write

$$(8.16) \quad u_n(x) = \tau_{n,l}(x) + r_{n,l}(x),$$

$$(8.16a) \quad \tau_{n,l}(x) = \int_{-l}^l \frac{1}{\lambda - \rho} d_\rho \int_D^{(m)} f(y) \theta_n(x, y | \rho) dT_y,$$

$$(8.16b) \quad r_{n,l}(x) = \left(\int_l^\infty + \int_{-\infty}^{-l} \right) \frac{1}{\lambda - \rho} d_\rho \int_D^{(m)} f(y) \theta_n(x, y | \rho) dT_y.$$

On making use of the symbol (8.14) and employing the extended definition (8.7) of θ_n one may express (8.16)–(8.16b) in the form

$$(8.17) \quad \int_{(C)}^{(x)} u_n(x) dT_x = \int_{(C)}^{(x)} \tau_{n,l}(x) dT_x + \int_{(C)}^{(x)} r_{n,l}(x) dT_x,$$

$$(8.17a) \quad \int_{(C)}^{(x)} \tau_{n,l}(x) dT_x = \int_{-l}^l \frac{1}{\lambda - \rho} d_\rho \int_{(C)}^{(x)} \int_C^{(m)} f(y) \theta_n(x, y | \rho) dT_x dT_y,$$

$$(8.17b) \quad \int_{(C)}^{(x)} r_{n,l}(x) dT_x = \left(\int_l^\infty + \int_{-\infty}^{-l} \right) \frac{1}{\lambda - \rho} d_\rho \int_{(C)}^{(x)} \int_C^{(m)} f(y) \theta_n(x, y | \rho) dT_x dT_y.$$

From (8.17a), (8.17b) we finally obtain

$$(8.18) \quad \int_{(C)}^{(x)} \tau_{n,l}(x) dT_x = \int_{-l}^l \frac{1}{\lambda - \rho} d_\rho \int_C^{(m)} \int_C^{(m)} b_x(t) f(y) \theta_n(t, y | \rho) dT_t dT_y,$$

$$(8.19) \quad \int_{(c)}^{(x)} \tau_{n,l}(x) dT_x = \left(\int_l^\infty + \int_{-\infty}^{-l} \right) \frac{1}{\lambda - \rho} d_\rho \int_c^{(m)} \int_c^{(m)} b_x(t) f(y) \theta_n(t, y | \rho) dT_t dT_y$$

for (x) in C ; here $b_x(t)$ is from (8.9).

Let T (Definition 6.1) be of the form O (cf. text subsequent to (6.14)). For λ in T one has

$$(8.20) \quad |\lambda - \lambda_{n,i}| \geq \delta = \delta(\lambda) > 0 \quad (n = k_1, k_2, \dots; i = 1, 2, \dots),$$

where $\delta(\lambda)$ is independent of n and i . Let $G(\delta)$ be the set of points whose distance from the interval $(-l, +l)$ is $< \delta$.

CASE A. The point represented by λ is on the frontier of $G(\delta)$ or is exterior $G(\delta)$.

CASE B. The point λ is in $G(\delta)$.

In Case A we have

$$(8.21) \quad |\lambda - \rho| \geq \delta > 0 \quad (\text{for all } \rho \text{ on } (-l, l))$$

and, consequently,

$$(8.21a) \quad \frac{1}{\lambda - \rho},$$

as a function of ρ , is continuous on the closed interval $(-l, l)$. In view of (7.25) from (8.18) it can be therefore deduced that

$$(8.22) \quad \lim_{k_j} \int_{(c)}^{(x)} \tau_{k_j,l}(x) dT_x = \int_{-l}^l \frac{1}{\lambda - \rho} d_\rho \int_c^{(m)} \int_c^{(m)} b_x(t) f(y) \theta(t, y | \rho) dT_t dT_y.$$

In Case B we take l sufficiently great so that the circle $S(\lambda, \delta)$, whose center is at λ and whose radius is δ , intersects the interval in two points, represented by numbers α, β , where

$$-l < \alpha < \beta < l.$$

The discussion, below, will still hold without any essential modifications when l is left unchanged. In view of (8.20) there are no points $\lambda_{n,i}$ ($n = k_1, k_2, \dots; i = 1, 2, \dots$) in the open interval (α, β) . Whence

$$\int_\alpha^\beta \frac{1}{\lambda - \rho} d_\rho \int_c^{(m)} \int_c^{(m)} b_x(t) f(y) \theta_n(t, y | \rho) dT_t dT_y = 0$$

($n = k_1, k_2, \dots$) and, by (8.18),

$$(8.23) \quad \int_{(c)}^{(x)} \tau_{n,l}(x) dT_x = \left(\int_{-l}^\alpha + \int_\beta^l \right) \frac{1}{\lambda - \rho} d_\rho \int_c^{(m)} \int_c^{(m)} b_x(t) f(y) \theta_n(t, y | \rho) dT_t dT_x.$$

In the field of integration indicated in (8.23) one has (8.21) and, accordingly, the function (8.21a) is continuous, in ρ , for ρ on the closed intervals

$$(-l, \alpha), \quad (\beta, l).$$

We apply (7.25) to each of the integrals

$$\int_{-l}^{\alpha} \dots, \quad \int_{\beta}^l \dots,$$

letting $n = k_j \rightarrow \infty$. From (8.23) it is thus inferred that

$$(8.24) \quad \lim_{k_j} \int_{(c)}^{(x)} \tau_{k_j, l}(x) dT_x = \left(\int_{-l}^{\alpha} + \int_{\beta}^l \right) \frac{1}{\lambda - \rho} \\ \cdot \int_c^{(m)} \int_c^{(m)} b_x(t) f(y) \theta(t, y | \rho) dT_t dT_y.$$

On the other hand, inasmuch as

$$\int_c^{(m)} \int_c^{(m)} b_x(t) f(y) \theta_n(t, y | \rho) dT_t dT_y = \gamma_n(x) \quad (n = k_1, k_2, \dots),$$

where $\gamma_n(x)$ is independent of ρ for $\alpha < \rho < \beta$, and in so far as the limit

$$(8.25) \quad \lim_{k_j} \int_c^{(m)} \int_c^{(m)} b_x(t) f(y) \theta_n(t, y | \rho) dT_t dT_y \\ = \int_c^{(m)} \int_c^{(m)} b_x(t) f(y) \theta(t, y | \rho) dT_t dT_y = \gamma(x)$$

exists (in view of the developments of section 7), it is concluded that $\gamma(x)$ of (8.25) is independent of ρ for $\alpha < \rho < \beta$. The implication of the latter fact is the relation

$$(8.26) \quad \int_{\alpha}^{\beta} \frac{1}{\lambda - \rho} d\rho \int_c^{(m)} \int_c^{(m)} b_x(t) f(y) \theta(t, y | \rho) dT_t dT_y = 0.$$

By virtue of (8.26) one may put (8.24) in the form of (8.22); that is, *the equality (8.22) holds in any case* (for every λ in T).

In order to study the function (8.19) we take l so great that

$$-l < R\lambda < l \quad (R\lambda = \text{real part of } \lambda).$$

It is then deduced that

$$(8.27) \quad |\lambda - \rho| \geq l - R\lambda > 0,$$

for $\rho \geq l$, and that

$$(8.27a) \quad |\lambda - \rho| \geq l + R\lambda > 0$$

for $\rho \leq -l$. In view of (8.27) we find that

$$\left| \int_l^\infty \frac{1}{\lambda - \rho} d_\rho \int_c^{(m)} \int_c^{(m)} b_z(t) f(y) \theta_{k_j}(t, y | \rho) dT_t dT_y \right| \\ \leq \frac{1}{l - R_\lambda} V_l^\infty \int_c^{(m)} \int_c^{(m)} b_z(t) f(y) \theta_{k_j}(t, y | \rho) dT_t dT_y$$

($j = 1, 2, \dots$). Now by virtue of (8.9) and (7.24)

$$V_l^\infty \dots \leq A = L^{\frac{1}{2}m} \left[\int_D^{(m)} |f(y)|^2 dT_y \right]^{\frac{1}{2}},$$

where L is from (8.6); A is independent of k_j . Thus

$$(8.28) \quad \left| \int_l^\infty \frac{1}{\lambda - \rho} d_\rho \int_c^{(m)} \int_c^{(m)} b_z(t) f(y) \theta_{k_j}(t, y | \rho) dT_t dT_y \right| \\ \leq A \cdot (l - R_\lambda)^{-1} \quad (j = 1, 2, \dots; (x) \text{ in } C).$$

Similarly, on making use of (8.27a) it is concluded that

$$(8.28a) \quad \left| \int_\infty^{-l} \text{integrand of the first member in (8.28)} \right| \leq \frac{A}{l + R_\lambda}$$

in C . Thus, by (8.28), (8.28a) and (8.19)

$$(8.29) \quad \left| \int_{(c)}^{(x)} r_{n,l}(x) dT_x \right| \leq A \left[\frac{1}{l - R_\lambda} + \frac{1}{l + R_\lambda} \right] \quad (\text{in } C)$$

for $n = k_1, k_2, \dots$ and for λ in T . For any λ , fixed as specified, the second member in (8.29) can be made arbitrarily small by suitable choice of l ; this can be done uniformly with respect to $n = k_j$ ($j = 1, 2, \dots$). Using this fact, as well as the relation (8.22) (valid in any case), in view of (8.17) it is deduced that

$$(8.30) \quad \lim_{k_j} \int_{(c)}^{(x)} u_{k_j}(x) dT_x = \int_{-\infty}^\infty \frac{1}{\lambda - \rho} d_\rho \int_c^{(m)} \int_c^{(m)} b_z(t) f(y) \theta(t, y | \rho) dT_t dT_y.$$

On taking account of (8.30), (8.9) and (8.13) it is inferred that

$$(8.31) \quad \int_{(c)}^{(x)} u(x) dT_x = \int_{-\infty}^\infty \frac{1}{\lambda - \rho} d_\rho \int_{(c)}^{(x)} \int_c^{(m)} f(y) \theta(x, y | \rho) dT_x dT_y.$$

Finally

$$(8.32) \quad u(x) = \frac{\partial^m}{\partial x_1 \dots \partial x_m} \int_{-\infty}^\infty \frac{1}{\lambda - \rho} d_\rho \int_{(c)}^{(x)} \int_D^{(m)} f(y) \theta(x, y | \rho) dT_x dT_y$$

for (x) in D .

In view of the above it is possible to formulate the following Theorem.

THEOREM 8.2. *Suppose D is bounded. Every solution $u(x)$, for λ in O , referred to in Theorem 6.2 and satisfying the non homogeneous problem (5.1), is expressible in the form (8.32) (cf. (8.14)).*

The results of this section can be extended to hold for D unbounded if the integration $\int_{(e)}^{(x)}$ is replaced by integration \int_e , where e are Borel measurable subsets of D , with frontiers of zero measure, and if the derivation $\frac{\partial^m}{\partial x_1 \dots \partial x_m}$ is replaced by set-function derivation D_x .

9. Some further developments involving spectra

We shall now essentially modify the procedures of section 8 and shall establish a number of spectral representations involving integrals convergent in the mean square. Throughout, \sim will denote convergence in the mean square. We shall first establish the following result.

THEOREM 9.1. *Suppose F is closed with respect to a spectrum θ (Definition 8.1). Then every function $a(x)$, $\subset L_2$ in D , has a following representation*

$$(9.1) \quad a(x) \sim \int_{-\infty}^{\infty} d\lambda \int_D^{(m)} \theta(x, y | \lambda) a(y) dT_y \quad ((x) \text{ in } D).$$

To prove this we note that by virtue of the last statement of section 7 there exists a function $b(x)$, $\subset L_2$ in D , for which

$$(9.2) \quad b(x) \sim \int_{-\infty}^{\infty} d\lambda \int_D^{(m)} \theta(x, y | \lambda) a(y) dT_y.$$

The function

$$\gamma(x) = b(x) - a(x)$$

belongs to L_2 ; in view of (8.3) and of the stated property of closure

$$(9.3) \quad \int_D^{(m)} \gamma^2(y) dT_y = \int_{-\infty}^{\infty} d\lambda \int_D^{(m)} \int_D^{(m)} \theta(x, y | \lambda) \gamma(x) \gamma(y) dT_x dT_y = \lim_l w_l,$$

where

$$(9.3a) \quad w_l = \int_{-l}^l d\lambda \int_D^{(m)} \int_D^{(m)} \theta(x, y | \lambda) \gamma(x) \gamma(y) dT_x dT_y.$$

Replacing here $\gamma(y)$ by $b(y) - a(y)$ and making use of the formula preceding (7.26) one obtains

$$(9.4) \quad w_l = w_{l,1} - w_{l,2},$$

where

$$w_{l,1} = \int_{-l}^l d\lambda \int_D^{(m)} \int_D^{(m)} \theta(x, y | \lambda) \gamma(x) b(y) dT_x dT_y$$

and

$$(9.5) \quad w_{l,2} = \int_{-l}^l d\lambda \int_D^{(m)} \int_D^{(m)} \theta(x, y | \lambda) \gamma(x) a(y) dT_x dT_y = \int_D^{(m)} \gamma(x) a_l(x) dT_x;$$

here

$$a_l(x) = \int_l^1 d_\lambda \int_D^{(m)} \theta(x, y | \lambda) a(y) dT_y.$$

In consequence of (8.5), which holds by virtue of the closure of F , one has

$$(9.6) \quad \lim_l w_{l,1} = \int_D^{(m)} \gamma(x) b(x) dT_x.$$

Now by (9.2)

$$\lim_l \int_D^{(m)} (b(x) - a_l(x))^2 dT_x = 0.$$

Hence from (9.5) it is inferred that

$$(9.6a) \quad \lim_l w_{l,2} = \int_D^{(m)} \gamma(x) b(x) dT_x.$$

From (9.4), (9.6), (9.6a) we finally conclude that

$$\lim_l w_l = 0.$$

Whence, as may be seen from (9.3), $\gamma(y) = 0$; thus $b(x) = a(x)$ and the relation (9.2) becomes the representation (9.1), which was to be established.

THEOREM 9.2. *Let λ be in O . The problem*

$$(9.7) \quad F(u) + \lambda u = f$$

has a solution ($\subset L_2$ in D)

$$(9.8) \quad u(x) \sim \int_{-\infty}^{\infty} \frac{1}{\lambda - \rho} d_\rho \int_D^{(m)} f(y) \theta(x, y | \rho) dT_y.$$

It will be sufficient to give the proof for real λ in O and for f real valued. We shall write

$$s_n(x, \rho) = \int_D^{(m)} f(y) \theta_n(x, y | \rho) dT_y, \quad s(x, \rho) = \int_D^{(m)} f(y) \theta(x, y | \rho) dT_y.$$

It will be first proved that, for $0 < l < +\infty$ and for λ in O ,

$$(9.9) \quad \lim_{n_i} \int_l^1 \frac{1}{\lambda - \rho} d_\rho s_{n_i}(x, \rho) = \int_l^1 \frac{1}{\lambda - \rho} d_\rho s(x | \rho).$$

With $\delta = \delta(\lambda)$ equal to the distance from λ to the closure of the set of points $\lambda_{n,i}$ ($n, i = 1, 2, \dots$) one has

$$d_\rho s_n(x | \rho) = 0$$

for

$$\lambda - \frac{\delta}{2} \leq \rho \leq \lambda + \frac{\delta}{2}$$

and

$$(9.10) \quad \int_{-l}^l \frac{1}{\lambda - \rho} d_\rho s_n(x | \rho) = \left(\int_{-l}^{\lambda'} + \int_{\lambda''}^l \right) \frac{1}{\lambda - \rho} d_\rho s_n(x | \rho)$$

$\left(\lambda' = \lambda - \frac{\delta}{2}, \lambda'' = \lambda + \frac{\delta}{2}; l \text{ suitably great} \right)$. By the formula preceding (7.25)

$$(9.10a) \quad \lim_{n_i} \int_{-l}^l \frac{1}{\lambda - \rho} d_\rho s_n(x | \rho) = \left(\int_{-l}^{\lambda'} + \int_{\lambda''}^l \right) \frac{1}{\lambda - \rho} d_\rho s(x | \rho)$$

$$= \int_{-l}^l \frac{1}{\lambda - \rho} d_\rho s(x | \rho),$$

inasmuch as

$$d_\rho s(x | \rho) = 0 \quad (\lambda' \leq \rho \leq \lambda'').$$

Thus (9.9) holds. We let

$$(9.11) \quad u_n(x, l) = \int_{-l}^l \frac{1}{\lambda - \rho} d_\rho s_n(x | \rho), \quad r_n(x, l) = \left(\int_{-\infty}^{-l} + \int_l^{\infty} \right) \frac{1}{\lambda - \rho} d_\rho s_n(x | \rho).$$

Then, in view of (8.15),

$$(9.11a) \quad u_n(x) = u_n(x, l) + r_n(x, l).$$

It has been established previously that the limit

$$\lim_{n_i} u_{n_i}(x) = u(x)$$

exists in D ; also,

$$(9.11b) \quad \lim_{n_i} u_{n_i}(x, l) = u(x, l) = \int_{-l}^l \frac{1}{\lambda - \rho} d_\rho s(x | \rho)$$

in consequence of (9.9). Hence, by (9.11a), the limit

$$(9.11c) \quad \lim_{n_i} r_{n_i}(x, l) = r(x, l)$$

similarly exists. With the notation so introduced one has

$$(9.11d) \quad u(x) = u(x, l) + r(x, l);$$

also

$$(9.12) \quad \int_D^{(m)} |u(x) - u(x, l)|^2 dT_x = \int_D^{(m)} |r(x, l)|^2 dT_x.$$

We have (admitting complex values), by virtue of (9.11),

$$|r_n(x, l)|^2 = \sum'_\alpha \sum'_\beta \frac{1}{\lambda - \lambda_{n,\alpha}} \frac{1}{\bar{\lambda} - \lambda_{n,\beta}} \int_D^{(m)} \int_D^{(m)} f(y) \bar{f}(z)$$

$$\cdot u_{n,\alpha}(x) u_{n,\beta}(x) u_{n,\alpha}(y) u_{n,\beta}(z) dT_y dT_z,$$

where the primes over the summation symbols indicate that the sums are taken

corresponding to the intervals $(-\infty, -l)$, $(l, +\infty)$. Using the orthonormality relations of the $u_{n,i}$ ($i = 1, 2, \dots$) one accordingly obtains

$$\int_D |r_n(x, l)|^2 dT_x = \sum'_\alpha \frac{1}{|\lambda - \lambda_{n,\alpha}|^2} \int_D^{(m)} \int_D^{(m)} f(y)\bar{f}(z) u_{n,\alpha}(y) \bar{u}_{n,\alpha}(z) dT_y dT_z.$$

The spectral form for this relation is

$$\begin{aligned} \int_D^{(m)} |r_n(x, l)|^2 dT_x &= \left(\int_{-\infty}^{-l} + \int_l^{\infty} \right) \frac{1}{|\lambda - \rho|^2} d_\rho \psi_n(\rho), \\ \psi_n(\rho) &= \int_D^{(m)} \int_D^{(m)} f(y)\bar{f}(z) \theta_n(y, z | \rho) dT_y dT_z. \end{aligned}$$

In consequence of (7.24)

$$V_\alpha^\beta \psi_n(\rho) \leq \int_D^{(m)} |f|^2 dT$$

for all finite α, β .

With $l > |R_\lambda|$ (λ real)

$$\left| \frac{1}{\lambda - \rho} \right| \leq \frac{1}{l - R_\lambda} \quad (\rho \geq l), \quad \left| \frac{1}{\lambda - \rho} \right| \leq \frac{1}{l + R_\lambda} \quad (\rho \leq -l).$$

Then

$$\int_D^{(m)} |r_n(x, l)|^2 dT_x \leq \left[\frac{1}{(l - R_\lambda)^2} + \frac{1}{(l + R_\lambda)^2} \right] \int_D^{(m)} |f|^2 dT = \sigma_\lambda(l).$$

The limiting function $r(x, l)$, of (9.11c), will satisfy the inequalities

$$\int_D^{(m)} |r(x, l)|^2 dT_x \leq \lim_n \int_D^{(m)} |r_n(x, l)|^2 dT_x \leq \sigma_\lambda(l).$$

Hence

$$\lim_{l \rightarrow \infty} \int_D^{(m)} |r(x, l)|^2 dT = 0.$$

Together with (9.12) and (9.11b), this implies that

$$\int_{-l}^l \frac{1}{\lambda - \rho} d_\rho s(x | \rho) \sim u(x) \quad (\text{in } D; \text{ as } l \rightarrow \infty),$$

which establishes (9.8) of the theorem.

We shall say that v is "admissible" if v is real, $\subset L_2$ (in D) and if

$$(9.13) \quad \frac{\partial v}{\partial \nu} + a_n(x)v = 0 \quad (\text{on } S_n; n = n_1, n_2, \dots)$$

for some functions $a_n(x)$, continuous on S_n ($n = 1, 2, \dots$).

It will be of interest to obtain a spectral representation for the differential operator F . We form the $u_{n,i}$, θ_n and θ , corresponding to the $a_n(x)$ ($n = n_1, n_2, \dots$) of (9.13). The following can be established (compare with (C)).

If v is "admissible" and $F(v) \subset L_2$ (in D), while F is closed with respect to θ , then

$$(9.14) \quad F(v(x)) \sim - \int_{-\infty}^{\infty} \lambda d\lambda \int_D^{(m)} \theta(x, y | \lambda) v(y) dT_y$$

(when there is only one spectrum, it will be independent of v , of course); moreover, in order that the integral of the second member in (9.14) should converge in the mean square it is necessary and sufficient that the integral

$$(9.15) \quad \int_{-\infty}^{\infty} \lambda^2 d\lambda \int_D^{(m)} \int_D^{(m)} \theta(x, y | \lambda) v(x) v(y) dT_x dT_y$$

should converge in the ordinary sense.

To prove (9.14) we apply the Fundamental Formula to $v, u_{n,i}$, obtaining the relation

$$(9.16) \quad \int_D^{(m)} F(v) u_{n,i} dT_x = -\lambda_{n,i} \int_D^{(m)} v u_{n,i} dT_x.$$

Now by Theorem 9.1

$$(9.17) \quad F(v(x)) \sim q_l(x) \quad (\text{as } l \rightarrow +\infty),$$

where

$$q_l(x) = \int_{-l}^l d\lambda \int_D^{(m)} \theta(x, y | \lambda) F(v(y)) dT_y.$$

In view of the formula preceding (7.25)

$$q_l(x) = \lim_n q_{l,n}(x) \quad (\text{as } n = n_i \rightarrow \infty),$$

with

$$q_{l,n}(x) = \int_{-l}^l d\lambda \int_D^{(m)} \theta_n(x, y | \lambda) F(v(y)) dT_y = \sum_i' u_{n,i}(x) \int_D^{(m)} F(v(y)) u_{n,i}(y) dT_y,$$

where the summation corresponds to the interval $(-l, l)$. By (9.16)

$$\begin{aligned} q_{l,n}(x) &= - \sum_i' \lambda_{n,i} u_{n,i}(x) \int_D^{(m)} v(y) u_{n,i}(y) dT_y \\ &= - \int_{-l}^l \lambda d\lambda \int_D^{(m)} \theta_n(x, y | \lambda) v(y) dT_y. \end{aligned}$$

In view of the relation preceding (7.25)

$$q_l(x) = - \int_{-l}^l \lambda d\lambda \int_D^{(m)} \theta(x, y | \lambda) v(y) dT_y.$$

This, together with (9.17), implies (9.14). The statement with respect to (9.15) can be proved with the aid of section 7.

10. Operators A and their application

Let λ be in O and let u_n be solution of the problem

$$F(u_n) + \lambda u_n = f \quad (\text{in } D_n),$$

$$\frac{\partial u_n}{\partial \nu} + a_n(x)u_n = 0 \quad (\text{on } S_n),$$

while w_n is solution of the problem

$$F(w_n) + \lambda w_n = q \quad (\text{in } D_n),$$

$$\frac{\partial w_n}{\partial \nu} + a_n(x)w_n = 0 \quad (\text{on } S_n);$$

here f, q are continuous and

$$(10.1) \quad f, q \in L_2 \quad (\text{in } D).$$

It is to be noted that the set O is independent of the choice of f and q . We define u_n, w_n as zero in $D - (D_n + S_n)$.

For $u = u_n$ and $v = w_n$ application of the Fundamental Formula (5.11) will yield

$$(10.2) \quad \int_D^{(m)} w_n f dT_x = \int_D^{(m)} u_n q dT_x \quad (u = n_1, n_2, \dots).$$

There exists a subsequence (k_1, k_2, \dots) of (n_1, n_2, \dots) such that

$$(10.3) \quad w_{k_j}(x) \rightarrow w(x), \quad u_{k_j}(x) \rightarrow u(x) \quad (\text{as } k_j \rightarrow \infty; \text{ in } D),$$

where

$$u(x), w(x) \in L_2 \quad (\text{in } D),$$

while, for (x) in D , one has

$$(10.4) \quad F(u) + \lambda u = f,$$

$$(10.4a) \quad F(w) + \lambda w = q;$$

moreover,

$$(10.5) \quad \int_D^{(m)} |w_{k_j}(x)|^2 dT_x, \quad \int_D^{(m)} |u_{k_j}(x)|^2 dT_x \leq \sigma(\lambda) < +\infty$$

($j = 1, 2, \dots$) with $\sigma(\lambda)$ independent of j . These facts are a consequence of the Existence Theorem 6.2.

In view of (10.5), (10.3) and (10.1) from (10.2) we derive

$$(10.6) \quad \int_D^{(m)} w f dT = \int_D^{(m)} u q dT.$$

By a reasoning of the type employed for an analogous purpose in (C) or using spectral representations one is able to choose the subsequence (k_1, k_2, \dots) independent of f and q .

The limits $u(x)$, $w(x)$ (cf. (10.3)), so obtained, can be represented as

$$(10.7) \quad u(x) = A_x(\lambda | f), \quad w(x) = A_x(\lambda | q).$$

THEOREM 10.1. *Let O be the set defined subsequent to (6.14). Suppose F is self adjoint. There exist an operator*

$$A_x(\lambda | \dots)$$

so that, whenever continuous $f, q \in L_2$ in D , formula (10.7) will represent solutions of equations (10.4), (10.4a), respectively, for λ in O . This operator satisfies the identity

$$(10.8) \quad \int_D^{(m)} A_x(\lambda | q) f(x) dT_x = \int_D^{(m)} A_x(\lambda | f) q(x) dT_x,$$

which is valid for every λ in O and for all f, q of the described type. $A_x(\lambda | \dots)$ may depend on the choice of the $a_n(x)$.

Theorem 10.1 is analogous to a result in (C_1 ; p. 56).

Let $f, q \in L_2$ (in D) and

$$(10.9); (10.10) \quad F(f), F(q) \in L_2 \quad (\text{in } D).$$

We then have

$$(10.11) \quad F(f) + \lambda f \equiv f^* \in L_2 \quad (\text{in } D),$$

$$(10.11a) \quad F(q) + \lambda q \equiv q^* \in L_2 \quad (\text{in } D).$$

Suppose F is of class I , with respect to O .

Application of (10.8) will enable inversion of (10.11), (10.11a) so as to yield

$$(10.12) \quad \int_D^{(m)} q(x) f^*(x) dT_x = \int_D^{(m)} f(x) q^*(x) dT_x.$$

Replacing f^*, q^* in (10.12) by the expressions from (10.11), (10.11a) one obtains

$$(10.13) \quad \int_D^{(m)} q(x) F(f) dT_x = \int_D^{(m)} f(x) F(q) dT_x.$$

The condition (10.13) is accordingly necessary in order that F be of Class I in O .

Conversely, suppose now that (10.13) holds for all functions $f, q \in L_2$ (in D) for which one has (10.10). If F were not of class I (for λ in O) there would be on hand a function $\varphi(x) \in L_2$ and a value λ_1 in O so that

$$(10.14) \quad F(\varphi) + \lambda_1 \varphi = 0,$$

$$(10.14a) \quad \int_D^{(m)} |\varphi(x)|^2 dT \neq 0.$$

Now, let us define $q(x)$ as a solution, $\in L_2$ (in D), of the equation

$$(10.15) \quad F(q) + \lambda_1 q = \varphi.$$

By Theorem 6.2 such a solution $q(x)$ exists. Having assumed (10.13), we shall have in particular

$$(10.16) \quad \int_D^{(m)} q(x)F(\varphi) dT_x = \int_D^{(m)} \varphi(x)F(q) dT_x.$$

Replacing $F(q)$ by the expression obtained from (10.15), in consequence of (10.16) one obtains

$$\int_D^{(m)} |\varphi(x)|^2 dT_x = \lambda_1 \int_D^{(m)} \varphi(x)q(x) dT_x + \int_D^{(m)} q(x)F(\varphi) dT_x.$$

Thus, by (10.14)

$$(10.17) \quad \int_D^{(m)} |\varphi(x)|^2 dT_x = \int_D^{(m)} q(x)[F(\varphi) + \lambda_1 \varphi] dT_x = 0.$$

Now (10.17) is contrary to (10.14a). Hence $F \subset I$ (in O). Accordingly the following Theorem has been established.

THEOREM 10.2. *Consider self adjoint operators F . In order that F be of class I (for λ in O), in accordance with Definition 6.2, it is necessary and sufficient that*

$$\int_D^{(m)} q(x)F(f) dT_x = \int_D^{(m)} f(x)F(q) dT_x$$

for all functions f, q belonging to L_2 (in D), for which $F(f), F(q)$ belong to L_2 (in D).

On repeatedly using (10.8), it can be shown that the following holds.

THEOREM 10.3. *If for a fixed non real λ_1 the equation*

$$(10.18) \quad F(u) + \lambda u = 0 \quad (\text{in } D)$$

has no solutions,²⁰ $\subset L_2$, the same will be true for all non real λ . For non real values of λ the number of distinct²¹ solutions, $\subset L_2$, of (10.18) is the same.

The proof of this theorem will be omitted, as it may be given following closely the lines of proof of analogous results in (C_1 ; pp. 55, 58); it would be necessary, however, to use some of the previous developments of this section.

We shall now consider questions of uniqueness of solutions for λ in O , that is, when λ may be real. Such considerations would correspond to certain developments given by Trjitzinsky⁴ for singular integral equations. The following will be proved.

THEOREM 10.4. 1°. *If for a fixed λ_1 , in O , the equation (10.18) has no solutions, $\subset L_2$, the same will be true for all non real λ .*

2°. *The number m of distinct solutions, $\subset L_2$, of (10.18) for any λ fixed in O is equal to or is greater than the number n of distinct solutions for non real values λ .*

Consider part 1°. If λ_1 is non real the conclusion 1° follows by virtue of

¹⁹ Apply theorem (C_1 ; pp. 132, 133).

²⁰ Here and in the sequel trivial, i.e. null, solutions, of homogeneous problems are disregarded.

²¹ I.e., linearly independent.

Theorem 10.3. Take λ_1 real in O .²² If 1° fails, there exists a non real value λ and a corresponding function $u, \in L_2$, so that (10.18) holds and

$$\int_D^{(m)} |u|^2 dT \neq 0.$$

One has

$$(10.19) \quad F(u) + \lambda_1 u = (\lambda_1 - \lambda)u.$$

Now the equation

$$F(\varphi) + \lambda_1 \varphi = f_1 \quad (f_1 \in L_2 \text{ in } D)$$

cannot have two distinct solutions, since otherwise their difference ω would satisfy (10.18) for λ_1 , while

$$\int_D^{(m)} |\omega|^2 dT \neq 0,$$

contrary to hypothesis. Thus, one may express the relation (10.19) in the form

$$(10.20) \quad u(x) = A_x(\lambda_1 | (\lambda_1 - \lambda)u).$$

Now by (10.19)

$$F(\bar{u}) + \lambda_1 \bar{u} = (\lambda_1 - \bar{\lambda})\bar{u}$$

and, in place of (10.20), one obtains

$$(10.20a) \quad \bar{u}(x) = A_x(\lambda_1 | (\lambda_1 - \bar{\lambda})\bar{u}).$$

In consequence of (10.8)

$$\int_D^{(m)} A_x(\lambda_1 | (\lambda_1 - \bar{\lambda})\bar{u})(\lambda_1 - \lambda)u dT_x = \int_D^{(m)} A_x(\lambda_1 | (\lambda_1 - \lambda)u)(\lambda_1 - \bar{\lambda})\bar{u} dT_x$$

and, substituting from (10.20), (10.20a), we deduce

$$(\lambda_1 - \lambda) \int_D^{(m)} |u|^2 dT_x = (\lambda_1 - \bar{\lambda}) \int_D^{(m)} |u|^2 dT_x$$

which implies that λ is real, contrary to our supposition.

We now proceed to part 2°. By Theorem 10.3 n is independent of λ . If 2° does not hold, there is a value λ , in O , for which the number m of distinct solutions,

$$u_1(x), \dots, u_m(x),$$

is less than n . By Theorem 10.3 λ must be real. Let λ^* be a non real value; with λ^* there will be associated n distinct solutions,

$$u_1^*(x), \dots, u_n^*(x).$$

We have

$$F(u_i) + \lambda u_i = 0, \quad F(u_i^*) + \lambda^* u_i^* = 0 \quad (\text{in } D).$$

²² Supposing that O has real points; cf. text subsequent to (6.14).

Since λ is real the u_i may be taken real. Any solution u^* for λ^* is expressible in the form

$$u^*(x) = \sum_1^n c_i u_i^*(x) \quad (\text{constants } c_i).$$

With $m < n$, the c_i may be chosen so that $u^*(x)$ will satisfy

$$(10.21) \quad \int_D^{(m)} |u^*(x)|^2 dT \neq 0, \quad \int_D^{(m)} u^*(x) \bar{u}(x) dT = 0$$

for all solutions $u(x)$ for the value λ .

We have

$$F(u^*) + \lambda^* u^* = 0, \quad F(\bar{u}^*) + \bar{\lambda}^* \bar{u}^* = 0;$$

whence

$$(10.22) \quad \begin{aligned} F(u^*) + \lambda u^* &= (\lambda - \lambda^*) u^* \equiv f, \\ F(\bar{u}^*) + \lambda \bar{u}^* &= (\lambda - \bar{\lambda}^*) \bar{u}^* \equiv \bar{f}. \end{aligned}$$

Accordingly

$$(10.23) \quad \begin{aligned} u^*(x) &= A_x(\lambda | f) + w_1(x), \\ \bar{u}^*(x) &= A_x(\lambda | \bar{f}) + w_2(x); \end{aligned}$$

here $w_1(x)$, $w_2(x)$ are some solutions of (10.18) for λ . Since the operator A can be so defined that

$$\bar{A}_x(\lambda | w) = A_x(\lambda | \bar{w}) \quad (\text{all } w \subset L_2),$$

we have

$$w_2(x) = \bar{w}_1(x).$$

By (10.8)

$$\int_D^{(m)} A_x(\lambda | \bar{f}) f dT_x = \int_D^{(m)} A_x(\lambda | f) \bar{f} dT_x.$$

Substitution of (10.23) and of the expression for f , \bar{f} from (10.22) will yield

$$(\lambda - \lambda^*) \int_D^{(m)} (\bar{u}^* - \bar{w}_1) u^* dT = (\lambda - \bar{\lambda}^*) \int_D^{(m)} (u^* - w_1) \bar{u}^* dT.$$

Further, in view of the property stated in connection with (10.21)

$$(\lambda - \lambda^*) \int_D^{(m)} |u^*|^2 dT_x = (\lambda - \bar{\lambda}^*) \int_D^{(m)} |u^*|^2 dT_x.$$

This leads to the conclusion that λ must be real; thus there arises a contradiction, which completes the proof of the Theorem. This can be extended to the case when the number of solutions could possibly be infinite.

By 1° of Theorem 10.4, if for a fixed λ_1 , in O , the equation (10.18) has no solutions, $\subset L_2$ and distinct from zero, F will be of class I for all non real λ ;

by the necessary part of Theorem 10.2 this would imply that the identity of that Theorem holds (for all f, q of stated type); in consequence of the sufficient part one then infers that $F \subset I$ in every set O . Hence we have the following: *if the equation*

$$F(u) + \lambda u = 0$$

has no solutions, $\subset L_2$ in D and distinct from zero, for a value λ_1 in some set O , then the same will be true for all λ in every conceivable set O .

In another paper the present author intends to obtain direct conditions on the $A_{i,j}, B_i, C$ under which F is of class I , the latter property being important in connection with questions of uniqueness and the closure of θ .

UNIVERSITY OF ILLINOIS

AND

INSTITUTE FOR ADVANCED STUDY

ON A THEOREM OF TANNAKA AND KREIN

By S. BOCHNER

(Received June 13, 1941)

Let G denote an arbitrary group, C the class of almost periodic functions on G and C_0 the sub-class of all *finite* linear combinations of (irreducible) representation coefficients. With the norm

$$\|f\| = \sup_x \epsilon_\sigma |f(x)|,$$

C is a Banach space, and C_0 is a dense but not closed subspace. Therefore a functional $L(f)$ which is additive, that is

$$(1) \quad L(af + bg) = aL(f) + bL(g), \quad \cdot$$

need not be bounded on C_0 and hence need not be continuable onto all of C . However if $L(f)$ is positive, that is

$$(2) \quad L(f) \geq 0 \text{ for } f \text{ (real and)} \geq 0, \quad f \in C_0$$

then $|L(g)| \leq 2\|g\| \cdot L(1)$ and thus $L(f)$ is bounded and has a (unique) extension onto all of C which is again positive. By a recent theorem of M. Krein,¹ assumption (2) can be replaced by the weaker assumption:

$$(3) \quad L(|g|^2) \geq 0 \text{ for } g \in C_0.$$

Thus, if $L(f)$ on C_0 has the properties (1) and (3) it also has property (2). This theorem of Krein includes an earlier theorem of T. Tannaka² that any functional on C_0 having property (1) and the additional property

$$L(f_1 \cdot f_2) = L(f_1) \cdot L(f_2)$$

has an extension onto C itself.

The proof of Krein is based on ideas of N. Wiener and I. Gelfand which are extraneous to the problem, and we are going to give a new proof which stays wholly within the technique of uniform approximation to elements of C by elements of C_0 .

We consider a complete set of irreducible representations

$$\{\varphi_{\rho q}(x)\}, \quad x \in G.$$

The letter ρ designates an element of an index set of suitable potency, and for

¹ On positive functionals on almost periodic functions, Doklady Moscou 30, pp. 9-12, (1941).

² Über den Dualitätssatz der nicht-kommutativen topologischen Gruppen, Tohoku J. Math., 45, pp. 1-12, (1938).

each $\rho, p, q = 1, \dots, h = h(\rho)$, where $h(\rho)$ is the (finite) dimension of the ρ^{th} representation. Any element of C_0 can be written in the form

$$(4) \quad f(x) = \sum_{\rho, p, q} a_{pq}^{\rho} \varphi_{pq}^{\rho}(x),$$

where only a finite number of the constants a_{pq}^{ρ} are $\neq 0$. Introducing

$$(5) \quad \gamma_{pq}^{\rho} = L(\varphi_{pq}^{\rho})$$

we have

$$L(f) = \sum_{\rho, p, q} a_{pq}^{\rho} \gamma_{pq}^{\rho}.$$

We call $L(f)$ a *special* functional if only a finite number of the constants (5) are $\neq 0$. In this case, since

$$|a_{pq}^{\rho}| \leq \|f\|,$$

we have

$$|L(f)| \leq \|f\| \cdot \sum_{\rho, p, q} |\gamma_{pq}^{\rho}|.$$

Thus $L(f)$ is bounded and has a continuous extension onto C . Furthermore every non-negative element of C is a uniform limit of squares of elements of C_0 . Hence we obtain

LEMMA. For any special $L(f)$, property (3) implies property (2).

We will next consider a *non-negative* almost periodic function $\Lambda(t)$ and its Fourier coefficients

$$\lambda_{pq}^{\rho} = M_t \{ \Lambda(t) \overline{\varphi_{pq}^{\rho}(t)} \}.$$

If the function $\Lambda(t)$ is a class function then

$$\lambda_{pq}^{\rho} = \lambda_{\rho} \delta_{pq}.$$

In this case, $\Lambda(t)$ is called a *weight-function*, and, if only a finite number of the values λ_{pq}^{ρ} are $\neq 0$, a *special* weight function.

Take any element (4) of C_0 which can be put in the form $f = |g|^2$, $g \in C_0$, and introduce the family of elements

$$f_y = f(xy^{-1}).$$

Since $f_y = |g(xy^{-1})|^2$, and

$$\varphi_{pq}^{\rho}(xy^{-1}) = \sum_{r=1}^h \varphi_{pr}^{\rho}(x) \overline{\varphi_{qr}^{\rho}(y)},$$

we see that (3) implies $L(f_y) \geq 0$ and therefore $M_t \{ \Lambda(t) L(f_t) \} \geq 0$, the latter inequality being explicitly

$$(6) \quad \sum_{\rho, p, q} a_{pq}^{\rho} \lambda_{\rho} \gamma_{pq}^{\rho} \geq 0.$$

Thus given the functional $L(f)$ and the weight functional $\Lambda(t)$ there exists a functional ΛL such that

$$\Lambda L(\varphi_{pq}^{\rho}) = \lambda_{\rho} \cdot \gamma_{pq}^{\rho}$$

and, what is decisive, if L has property (3) then so does ΛL . Now if $\Lambda(t)$ is a special weight function then ΛL is a special functional, and by our lemma, (6) holds for any $f \geq 0$ belonging to C_0 .

Now let f be any *fixed* non-negative element from C_0 , and consider a sequence of weight functions, then

$$\sum_{\rho, p, q} a_{pq}^{\rho} \lambda_{\rho}^{(n)} \gamma_{pq}^{\rho} \geq 0.$$

where ρ varies over a fixed *finite* index set, say $\rho = 1, \dots, m$, which is independent of n , and $n = 1, 2, \dots$. However, given any *finite* index set $\{\rho\}$ there exists a sequence of special weight functions such that for each ρ from the set, $\lim_{n \rightarrow \infty} \lambda_{\rho}^{(n)} = 1$.³ Hence we obtain

$$\sum_{\rho, p, q} a_{pq}^{\rho} \gamma_{pq}^{\rho} \geq 0$$

for each non-negative element of C_0 , and this completes the proof of (2).

PRINCETON UNIVERSITY

³ S. Bochner and J. von Neumann, *Almost periodic functions in groups*, Trans. Amer. Math. Soc., 37, pp. 21-50 (1935), especially pp. 37-40.

ON THE UNIFORM DISTRIBUTION OF THE ROOTS OF CERTAIN POLYNOMIALS

BY P. ERDÖS

(Received July 21, 1941)

Let

$$\begin{array}{ccccccc} & & & & x_1^{(1)} & & \\ & & & & & & \\ & & x_1^{(2)} & & x_2^{(2)} & & \\ & & \dots & \dots & \dots & \dots & \\ x_1^{(n)} & & x_2^{(n)} & \dots & x_n^{(n)} & & \end{array}$$

be a triangular matrix, where, for each n ,

$$1 \geq x_1^{(n)} > x_2^{(n)} > \dots > x_n^{(n)} \geq -1.$$

Since $x_i^{(n)}$ may be written in the form $x_i^{(n)} = \cos(\vartheta_i^{(n)})$, where $0 \leq \vartheta_i^{(n)} \leq \pi$, we may define another triangular matrix

$$\begin{array}{ccccccc} & & & & \vartheta_1^{(1)} & & \\ & & & & & & \\ & & \vartheta_1^{(2)} & & \vartheta_2^{(2)} & & \\ & & \dots & \dots & \dots & \dots & \\ \vartheta_1^{(n)} & & \vartheta_2^{(n)} & \dots & \vartheta_n^{(n)} & & \end{array}$$

with

$$0 \leq \vartheta_1^{(n)} < \vartheta_2^{(n)} < \dots < \vartheta_n^{(n)} \leq \pi.$$

Put $\omega_n(x) = \prod (x - x_i)$.¹ Suppose $0 \leq A < B \leq \pi$. We denote by $N_n(A, B)$ the number of the ϑ_i in (A, B) . Let $-1 \leq a < b \leq 1$. Then we denote by $M_n(a, b)$ the number of the x_i in (a, b) . It does not matter whether the intervals (A, B) and (a, b) are open or closed.

In a previous paper² Turán and the author proved that if

$$|\omega_n(x)| < \frac{f(n)}{2^n}$$

then

$$N_n(A, B) = \frac{B - A}{\pi} n + O(n^{\frac{1}{2}}(\log f(n))^{\frac{1}{2}}).$$

¹ We omit the upper index n where there is no danger of confusion.

² On the uniformly dense distribution of certain sequences of points, *Annals of Math.* Vol. 41 (1940), pp. 162-173.

In another paper³ we proved that if $|l_k^{(n)}(x)| < c_1$ then

$$N_n(A, B) = \frac{B-A}{\pi} n + O[(B-A)n]^{1+\epsilon}.$$

($l_k^{(n)}(x)$ denotes the fundamental polynomials, i.e. $l_k^{(n)}(x) = \omega(x)/[\omega'(x_k)(x-x_k)]$ is of degree $n-1$, and $l_k(x_k) = 1$, $l_k(x_i) = 0$, $i \neq k$.)

In the present paper we are going to improve these results. First we prove

THEOREM 1. Put $x_0 = -1$, $x_{n+1} = 1$, and let

$$(1) \quad \max_{-1 \leq x \leq 1} |\omega_n(x)| < \frac{c_2}{2^n} \quad \text{and} \quad \max_{x_k \leq x \leq x_{k+1}} |\omega_n(x)| > \frac{c_3}{2^n}, \quad k = 0, 1, \dots, n.$$

Then

$$N_n(A, B) = \frac{B-A}{\pi} n + O[\log n(B-A)].$$

This result is the best possible.

Next we prove

THEOREM 2. Let $|l_k(x)| < c_4$; then

$$N_n(A, B) = \frac{B-A}{\pi} n + O[(\log n)(\log n(B-A))]$$

if $|l_k(x)| < n^{c_5}$, then

$$N_n(A, B) = \frac{B-A}{\pi} n + O[(\log n)^2].$$

Theorem 2 is also the best possible. Theorems 1 and 2 can be generalized to

THEOREM 3. Let $\omega(x)$ be such that

$$\frac{c_6 f(n)}{2^n} < \max_{x_k < x \leq x_{k+1}} |\omega_n(x)| < \frac{c_7 f(n)}{2^n} \quad k = 0, 1, 2, \dots, n;$$

then

$$N_n(A, B) = \frac{B-A}{\pi} n + O[(\log n)(\log f(n))].$$

Similarly, if $|l_k(x)| < c_8 f(n)$ then

$$N_n(A, B) = \frac{B-A}{\pi} n + O[(\log n)(\log n f(n))].$$

To prove Theorem 1 we first have to prove two lemmas.

LEMMA 1. Suppose that (1) holds; then

$$(2) \quad \frac{c_9}{n} < \vartheta_{k+1} - \vartheta_k < \frac{c_{10}}{n}, \quad k = 0, 1, \dots, n.$$

³ On interpolation iii, *ibid.* pp. 510-553.

PROOF. A theorem of M. Riesz states that if $h(x)$ is a polynomial of degree n which assumes its absolute maximum in $(-1, 1)$ at the point x_0 , and if y_1, \dots, y_r are the roots of $h(y) = 0$ in the interval $(-1, 1)$, then $|\theta_i - \theta_0| \geq \pi/2n$, where $\cos \theta_1 = y_i$ and $\cos \theta_0 = x_0$. Thus if x_0 lies between the roots y_i and y_{i+1} , then $\theta_{i+1} - \theta_i \geq \pi/n$. Also if $\max_{y_i \leq x \leq y_{i+1}} h(x)$ assumes its smallest value for $i = k$ then

$$\theta_{k+1} - \theta_k \leq \pi/n.$$

Suppose that (2) does not hold, for example assume that

$$\vartheta_{k+1} - \vartheta_k > r(n)/n,$$

where $\lim_{n \rightarrow \infty} r(n) = \infty$. Take $\epsilon > 0$, and define u and v by the relations: u and v are symmetric with respect to $(x_k + x_{k+1})/2$, and $\arccos u - \arccos v = \pi/n + \epsilon$. Consider the polynomial $\phi(x) = \omega(x) \cdot (x - u) \cdot (x - v) / (x - x_k)(x - x_{k+1})$. It can be seen that if $u \leq x \leq v$ then

$$\frac{(x - u)(x - v)}{(x - x_k)(x - x_{k+1})} < c_{11}/r(n);$$

hence

$$(3) \quad \max_{u \leq x \leq v} |\phi(x)| < (c_{11}/r(n)) \max_{x_k \leq x \leq x_{k+1}} \omega_n(x).$$

Also, since the sum of two quantities whose sum is fixed increases as they tend to equality, we have, in the intervals $(-1, x_k)$ and $(x_{k+1}, 1)$,

$$(4) \quad |\phi(x)| > |\omega(x)|.$$

We have $\arccos v - \arccos u > \pi/n$; and a simple calculation shows that, if $r(n)$ is large enough, $\vartheta_{k+1} - \arccos v > \pi/n$ and $\arccos u - \vartheta_k > \pi/n$; thus it follows from the lemma of M. Riesz (applied to $\phi(x)$) that $\max |\phi(x)|$ between two consecutive roots of $\phi(x)$, assumes its smallest value between the roots x_i and x_{i+1} , where either $i \leq k - 2$ or $i \geq k + 2$. Thus, from (3) and (4),

$$\max_{x_k \leq x \leq x_{k+1}} |\omega(x)| > \frac{r(n)}{c_{11}} \cdot \min_{i=0,1,\dots,n} \max_{x_i \leq x \leq x_{i+1}} |\omega(x)|.$$

This contradicts (1), which completes the proof. By the same argument we could prove the other inequality in (2).

COROLLARY. We obtain from Lemma 1, by a simple computation, that

$$\frac{c_{12}}{n} \cdot (1 - x_k^2)^{\frac{1}{2}} < x_{k+1} - x_k < \frac{c_{13}}{n} (1 - x_k^2)^{\frac{1}{2}}, \quad (k = 1, 2, \dots, n - 1).$$

LEMMA 2. Suppose that (1) holds; then for $-1 \leq x \leq 1$,

$$|l_k(x)| < c_{14} \frac{(1 - x_k^2)^{\frac{1}{2}}}{(x - x_k)n}.$$

PROOF. We have $l_k(x) = \omega(x)/[\omega'(x_k)(x - x_k)]$; thus by (1) it suffices to show that

$$\omega'(x_k) > c_{15} \frac{n}{2^n(1 - x_k^2)^{\frac{1}{2}}}.$$

Consider the polynomial $\psi(x) = \omega(x)/(x - x_k)$. It is clear that either $|\omega'(x_k)| = \psi(x_k) \geq |\psi(y)|$ if $x_{k-1} \leq y \leq x_k$, or $|\omega'(x_k)| = \psi(x_k) \geq \psi(y)$ if $x_k \leq y \leq x_{k+1}$. Without loss of generality we can assume that the first inequality holds. Then by (1) and the corollary to lemma 1 we have

$$\omega'(x_k) = \frac{\max_{x_{k-1} \leq x \leq x_k} |\omega(x)|}{x_k - x_{k-1}} > c_{15} \frac{n}{2^n(1 - x_k^2)^{\frac{1}{2}}},$$

which completes the proof.

Now we can prove Theorem 1. To simplify the calculations we assume that $a = 0$, $b = 1$. Then we have to show that, assuming (1)

$$\frac{n}{2} - c_{15} \log n < M_n(0, 1) < \frac{n}{2} + c_{17} \log n.$$

It will be sufficient to prove the first inequality. Suppose that it does not hold; then

$$M_n(0, 1) < \frac{n}{2} - r(n) \log n, \quad \overline{\lim} r(n) = \infty.$$

Consider the polynomial $g(x)$ whose roots are defined as follows: In the interval $(-1, \log n/n)$, $g(x)$ has the same roots as $T_{n-1}(x)$ ($T_n(x)$ denotes the n^{th} Tchebicheff polynomial); at the points $(\frac{3}{2})^r \log n/n$, $r = 1, 2, \dots, s$ where s is such that

$$\left(\frac{3}{2}\right)^s \frac{\log n}{n} \leq 1 < \left(\frac{3}{2}\right)^{s+1} \frac{\log n}{n},$$

$g(x)$ has a root of multiplicity $\left[\frac{r(n)}{10}\right]$; and finally $g(x)$ vanishes at the roots of $\omega(x)$ in the interval $(0, 1)$. Clearly the degree of $g(x)$ does not exceed

$$\frac{n}{2} + \log n + \frac{3 \log n}{10} r(n) + \frac{n}{2} - r(n) \log n < n - 1$$

if $r(n) > 10$. Thus, by the lemma of M. Riesz, $g(x)$ assumes its absolute maximum in the interval $(\log n/n, 1)$. Suppose that it assumes its absolute maximum at x_0 , $\log n/n \leq x_0 \leq 1$. We have for some r

$$\left(\frac{3}{2}\right)^r \frac{\log n}{n} < x_0 < \left(\frac{3}{2}\right)^{r+1} \frac{\log n}{n}.$$

(If $(\frac{3}{2})^{r+1} \log n/n > 1$, we replace it by 1.) Put $(\frac{3}{2})^r \log n/n = q$; we consider the polynomial

$$g_1(x) = \frac{g(x)}{(x - q)^p}, \quad p = \left[\frac{r(n)}{10} \right].$$

By the Lagrange interpolation formula we evidently have

$$g_1(x) = \sum_{k=1}^n g_1(x_k) l_k(x)$$

where the x_k are the roots of $\omega(x)$. Thus

$$(5) \quad g_1(x_0) = \sum_{k=1}^n g_1(x_k) l_k(x_0).$$

Now $g_1(x_k) = 0$ for $0 \leq x \leq 1$; and since x_0 was the place where $g(x)$ takes its absolute maximum, we have

$$g_1(x_0) \geq [2(t + 1)]^p g_1(x)$$

if x satisfies

$$(6) \quad -\frac{\log n}{n} t \left(\frac{3}{2} \right)^r \geq x \geq -\frac{\log n}{n} (t + 1) \left(\frac{3}{2} \right)^r; \quad t = 0, 1, 2 \dots$$

(6) may be verified by noting that

$$g_1(x_0) = \frac{g(x_0)}{(x_0 - q)^p} \geq \frac{g(x)}{(x_0 - q)^p} = g_1(x) \left(\frac{x - q}{x_0 - q} \right)^p.$$

Hence from (5) and (6), by putting

$$-\frac{\log n}{n} t \left(\frac{3}{2} \right)^r = u_t,$$

we obtain

$$1 \leq \sum_{t \geq 0} \frac{M_n(u_t, u_{t+1}) \max_{u_t \leq x_k \leq u_{t+1}} |l_k(x_0)|}{2(t + 1)^p} = \sum_1 + \sum_2$$

where in \sum_1 t is restricted by $u_t \geq -\frac{1}{2}$. Now by the corollary to Lemma 1, and Lemma 2.

$$\sum_1 < \sum_{t \geq 0} c_{18} n(u_{t+1} - u_t) c_{19} \frac{1}{n_{t+1}} \frac{1}{[2(t + 1)]^p} < c_{20} \sum_{t \geq 0} \frac{1}{[2(t + 1)]^p} < \frac{1}{2}$$

for sufficiently large p .

For the x_k in \sum_2 we clearly have $x_k < -\frac{1}{2}$. Thus by lemma 2.

$$\sum_2 < c_{21} \frac{n}{2^p} \max_{x_k < -\frac{1}{2}} |l_k(x_0)| < \frac{1}{2}$$

for sufficiently large p . Thus $\sum_1 + \sum_2 < 1$, and this contradiction establishes the proof.

In the proof we did not use the full strength of Lemma 2; in fact we only used

$|l_k(x)| < c_{22} \frac{1}{n|x-x_k|}$. We would have had to use the sharper estimate if we had not restricted ourselves to the interval $(0, 1)$ but had considered a "small" interval near -1 or $+1$.

Now we have to prove that the error term in Theorem 1 is the best possible. Put

$$\vartheta_0 = \frac{\pi}{2}, \quad \vartheta_k = \frac{\pi}{2} + \frac{k\pi}{n} + \sum_{i=1}^k \frac{1}{i}, \quad \vartheta_l = \frac{\pi}{2} - \frac{l\pi}{n} - \sum_{i=1}^l \frac{1}{i}$$

where k and l take all positive integral values such that $\vartheta_k < \pi - n^{-2}$, and $\vartheta_l > n^{-2}$ it is easy to see that the number of the ϑ 's is $n + O(1)$. Consider the polynomial $\omega(x)$ whose roots are the $\cos \vartheta$'s. It can be shown by elementary computations that $\omega(x)$ satisfies (1). We do not give the details. On the other hand it is easy to see that

$$M_n(0, 1) < \frac{n}{2} - c_{23} \log n$$

which shows that the error term in Theorem 1 is the best possible.

The proof of Theorem 2 is very similar to that of Theorem 1. The difference is that, in defining $g(x)$, $g(x)$ now has roots of order $\left[\frac{r(n) \log n}{10} \right]$ at the points $(\frac{3}{2})^r \log n/n$. The proof of Theorem 3 also runs along the same lines.

UNIVERSITY OF PENNSYLVANIA

ON THE ASYMPTOTIC DENSITY OF THE SUM OF TWO SEQUENCES

By P. ERDÖS

(Received June 24, 1941)

Let $a_1 < a_2 < \dots$ be an infinite sequence, A , of positive integers. Denote the number of a 's not exceeding n by $f(n)$. Schnirelmann has defined the density of A as G.L.B. $f(n)/n$.¹ Now let $a_1 < a_2 < \dots$; $b_1 < b_2 < \dots$ be two sequences. We define the sum $A + B$ of these two sequences as the set of integers of the form a_i or b_j or $\{a_i + b_j\}$. Schnirelmann proved that if the density of A is α and that of B is β then the density of $A + B$ is $\geq \alpha + \beta - \alpha\beta$.

Khintchine² proved that, provided that $\alpha = \beta \leq \frac{1}{2}$, the density of $A + B$ is $\geq 2\alpha$. He conjectured more generally that if $\alpha + \beta \leq 1$ the density of $A + B$ is $\geq \alpha + \beta$. It is easy to see that if $\alpha + \beta \geq 1$ then every integer is in $A + B$, so the density of $A + B$ is 1. Khintchine's conjecture seems very deep.

Besicovitch³ defined $\beta' = \text{G.L.B. } \varphi(n)/(n+1)$ where $\varphi(n)$ denotes the number of the b 's not exceeding n , and proved that the Schnirelmann density of the sequence of numbers $\{a_i, a_i + b_j\}$ is $\geq \alpha + \beta'$. An example of Rado showed that this result is the best possible.

Define the asymptotic density of A as $\lim f(n)/n$. Then if $\alpha \leq \frac{1}{2}$ and $a_1 = 1$ I have proved that the asymptotic density of $A + B$ is $\geq \frac{3}{2}\alpha$.⁴ The following simple example of Heilbronn shows that this result is the best possible: Let the a 's be the integers $\equiv 0, 1 \pmod{4}$. Then $A + A$ contains the integers $\equiv 0, 1, 2 \pmod{4}$. In the present note we prove the following

THEOREM: *Let the asymptotic density of A be α and that of B be β , where $\alpha + \beta \leq 1$, $\beta \leq \alpha$, $b_1 = 1$. Then the asymptotic density of $A + B$ is not less than $\alpha + \frac{1}{2}\beta$, and, in fact, one of the sequences $\{a_i, a_i + 1\}$ or $\{a_i + b_j\}$ has asymptotic density $\geq \alpha + \frac{1}{2}\beta$.*

It is easy to see that if $\alpha + \beta > 1$ then all large integers are in $A + B$. For if not then, none of the integers $n - a_i$ belong to B , and the asymptotic density of B would be not greater than $1 - \alpha < \beta$.

To prove our theorem we first need a slight sharpening of the theorem of Besicovitch; in fact, we prove the following

LEMMA: *Define the modified density of B as follows:*

¹ Schnirelmann, *Über additive Eigenschaften der Zahlen*, Math. Annalen 107 (1933), pp. 649-690.

² Khintchine, *Zur additiven Zahlentheorie*, Recueil math. de la soc. Moscow 39 (1932), pp. 27-34.

³ Besicovitch, *On the density of the sum of two sequences of integers*, Journ. of the London math. soc. 10 (1935), pp. 246-248.

⁴ Erdős, *On the asymptotic density of the sum of two sequences one of which forms a basis for the integers. ii.*, Travaux de l'institut math. de Tblissi 3 (1938), pp. 217-223.

$$1) \quad \beta_1 = \text{G.L.B.}_{n > k} \frac{\varphi(n)}{n+1},$$

where the integers $1, 2, \dots, k$ belong to B , but $k+1$ does not belong to B . Clearly $\beta_1 \geq \beta'$. Then the Schnirelmann density of the sequence $\{a_i, a_i + b_j\}$ is not less than $\alpha + \beta_1$.

The proof of this lemma follows closely the proof of Besicovitch. Denote by $f(u, v)$, $\varphi(u, v)$, $\psi(u, v)$ respectively the number of a 's, b 's, and terms of the sequence $\{a_i, a_i + b_j\}$ in the interval (u, v) —that is, among the integers $u+1, u+2, \dots, v$. We first observe that if $r+1$ is any integer which does not belong to the sequence $\{a_i, a_i + b_j\}$ then

$$2) \quad f(u, v) + \varphi(r-v, r-u) \leq v-u.$$

For as t runs through (u, v) , $r+1-t$ runs through $(r-v, r-u)$, and if t belongs to A then $r+1-t$ does not belong to B .

We may assume that the Schnirelmann density of the sequence $\{a_i, a_i + b_j\}$ is less than 1, and that $\alpha > 0$, so that $a_1 = 1$. Define $m_0 = 0$, define $r_0 + 1$ as the least positive integer not belonging to $\{a_i, a_i + b_j\}$, define $m_1 + 1$ as the least integer greater than r_0 belonging to A , define $r_1 + 1$ as the least integer greater than m_1 not belonging to $\{a_i, a_i + b_j\}$, and so on.

It suffices to prove that for each x in (r_{i-1}, m_i) we have

$$3) \quad \psi(0, x) \geq (\alpha + \beta_1)x,$$

for if (3) holds, suppose that for some y in (m_j, r_j) we had

$$\psi(0, y) < (\alpha + \beta_1)y.$$

(We may suppose $j > 0$; else $y \leq r_0$, so that $\psi(0, y) = y$). Then since all the integers $m_j + 1, \dots, y$ belong to $\{a_i, a_i + b_j\}$ and $\alpha + \beta_1 \leq 1$ we should have

$$\psi(m_j) < (\alpha + \beta_1)m_j,$$

which contradicts (3).

It follows from the definition of k and the definition of m_i and r_i that

$$4) \quad r_i - m_i > k \quad (i = 0, 1, 2, \dots).$$

Let $r_{i-1} < x \leq m_i$; we have

$$5) \quad \psi(r_{i-1}, x) \geq \varphi(r_{i-1} - m_{i-1} - 1, x - m_{i-1} - 1),$$

since any number $m_{i-1} + 1 + u$, where u belongs to B , is in $\{a_i, a_i + b_j\}$. Also

$$6) \quad \psi(m_{i-1}, r_{i-1}) = r_{i-1} - m_{i-1} \geq f(m_{i-1}, r_{i-1}) + \varphi(0, r_{i-1} - m_{i-1})$$

by (2). Clearly by the definition of the numbers r_i, m_i we have for $r_{i-1} < x \leq m_i$, $f(m_{i-1}, x) = f(m_{i-1}, r_{i-1})$. Hence by adding (5) and (6)

$$7) \quad \psi(m_{i-1}, x) \geq f(m_{i-1}, x) + \varphi(0, x - m_{i-1} - 1) \geq f(m_{i-1}, x) + \beta_1(x - m_{i-1}),$$

since by (4) $x - m_{i-1} - 1 \geq r_{i-1} - m_{i-1} > k$. In particular

$$8) \quad \psi(m_i, m_{i+1}) \geq f(m_i, m_{i+1}) + \beta_1(m_{i+1} - m_i) \quad (j = 0, 1, \dots).$$

Summing (8) for $j = 0, 1, \dots, i-1$ and adding (7) we have

$$\psi(0, x) \geq f(0, x) + \beta_1 x \geq (\alpha + \beta_1)x,$$

which completes the proof of the Lemma.

Now we can prove our theorem. We may assume $\beta > 0$. Suppose first that there exists an x belonging to A , such that the modified density of (the positive terms of) $a_i - x$ is $\geq \alpha - \frac{1}{2}\beta$. Clearly $x + 1$ has to be in A since $\alpha - \frac{1}{2}\beta > 0$. It follows that there exists for every positive real ϵ a y such that the Schnirelmann density of the positive terms of the sequence $\{b_j - y\}$ is $\geq \beta - \epsilon$. To see this choose y to be the greatest integer with

$$\frac{\varphi(y)}{y} \leq \beta - \epsilon.$$

(Since $\liminf \varphi(y)/y = \beta$ such a y exists, unless $\varphi(y)/y > \beta - \epsilon$ for all positive y ; in this case we have $y = 0$). Then by the definition of y it is clear that $\varphi(y, z)$ i.e. the number of $\{b_j - y\}$'s in $(0, z - y)$, is not less than $(\beta - \epsilon)(z - y)$, which proves our assertion.

Now consider the sequence $\{b_j - y, b_j - y + a_i - x\}$. By our lemma its Schnirelmann density is $\geq \alpha + \frac{1}{2}\beta - \epsilon$; hence by adding $x + y$ to its members we obtain the sequence $\{b_j + x, a_i + b_j\}$ whose asymptotic density is clearly $\geq \alpha + \frac{1}{2}\beta - \epsilon$ for every $\epsilon > 0$. But since x is in A , $b_j + x$ is in $\{a_i + b_j\}$. Hence the asymptotic density of the sequence $\{a_i + b_j\}$ is $\geq \alpha + \frac{1}{2}\beta$, which proves our theorem in the first case.

Suppose next that Case 1 is not satisfied. We may suppose that there exist arbitrarily large values of i such that a_i and $a_i + 1$ are both in A ; otherwise $\{a_i, a_i + 1\}$ has asymptotic density $2\alpha > \alpha + \frac{1}{2}\beta$. Let a_{k_1} be the first a_i such that $a_{k_1} + 1$ is also in A . Then since Case 1 is not satisfied and since $\alpha = \liminf f(n)/n$, there exists a largest integer m_1 such that $f(a_{k_1}, m_1) < (\alpha - \frac{1}{2}\beta)(m_1 - a_{k_1} + 1)$. Again let a_{k_2} be the least a_i greater than m_1 such that $a_{k_2} + 1$ is also in A ; there exists as before a largest m_2 such that $f(a_{k_2}, m_2) < (\alpha - \frac{1}{2}\beta)(m_2 - a_{k_2} + 1)$ and so on. Take n large and let m_r be the least $m \geq n$. It is clear that the intervals $(a_{k_i} - 1, m_i)$, $i = 1, 2, \dots, r$ do not overlap; thus

$$\sum_{i=1}^r f(a_{k_i}, m_i) \leq m_r \left(\alpha - \frac{\beta}{2} \right).$$

Now since the asymptotic density of A is α , we have $f(0, m_r) > (\alpha - \epsilon)m_r$, if n is large enough, and therefore the number of a_i 's in $(0, n)$ outside the intervals (a_{k_i}, m_i) , $i = 1, 2, \dots, r$ is not less than

$$\left(\frac{\beta}{2} - \epsilon \right) m_r \geq \left(\frac{\beta}{2} - \epsilon \right) n.$$

But for all these a_i 's with the exception of $a_{k_1}, a_{k_2}, \dots, a_{k_r}, a + 1$ is not in A .

Moreover, the intervals (a_{k_i}, m_i) do not contain only a 's; else, whenever $p > a_{k_i}$ is such that (a_{k_i}, p) does contain integers not in A , we have $p > m_i$. Therefore $f(a_{k_i}, p) \geq (\alpha - \frac{1}{2}\beta)(p - a_{k_i} + 1)$ (by definition of m_i); so that the modified density of the positive terms of $\{a_j - a_{k_i}\}$ ($j = 1, 2, \dots$) is $\geq \alpha - \frac{1}{2}\beta$, and we are in Case 1. Thus each of the intervals (a_{k_i}, m_i) has to contain an x which is in A , such that $x + 1$ is not in A . Hence, finally, the number of integers $\leq n$ of the form $a_i + 1$ which are not in A is $\geq (\frac{1}{2}\beta - \epsilon)(n - 1)$. Hence the number of integers $\leq n$ of the form $\{a_i, a_i + 1\}$ is not less than $(\alpha + \frac{1}{2}\beta - \epsilon)n - 1$, if n is large enough, which completes the proof of our theorem.

UNIVERSITY OF PENNSYLVANIA

SOME NEW SUMMABILITY METHODS WITH APPLICATIONS

BY OTTO SZÁSZ¹

(Received November 4, 1941)

1. Given an infinite sequence of functions

$$(1.1) \quad \varphi_0(x), \varphi_1(x), \varphi_2(x), \dots, \varphi_\nu(x), \dots,$$

defined on a finite or infinite range R of the real or complex variable x with the limit point ξ ; the range R may be discrete or continuous, but must contain infinitely many points. The *series-to-function transform*

$$(1.2) \quad \Phi(x) = \sum_{\nu=0}^{\infty} u_\nu \varphi_\nu(x)$$

defines, under certain assumptions for (1.1), a summability method, and $\lim_{x \rightarrow \xi} \Phi(x)$, if it exists, is called the generalized sum of the series $\sum_0^\infty u_\nu$. If it is true that for every convergent series $\sum u_\nu$, $\Phi(x)$ exists in R and $\lim \Phi(x) = \sum u_\nu$, then the method is called regular. The necessary and sufficient conditions for regularity are:

$$(1.3) \quad \lim_{x \rightarrow \xi} \varphi_\nu(x) = 1, \quad \text{for } \nu = 0, 1, 2, \dots$$

$$(1.4) \quad \sum_0^\infty |\varphi_\nu(x) - \varphi_{\nu+1}(x)| \text{ uniformly bounded on } R.$$

We then say that (1.2) is a regular transform; the summability method defined by

$$(1.5) \quad \lim_{x \rightarrow \xi} \Phi(x) = s$$

is called the method of convergence factors (according to C. N. Moore).² Corresponding to this method we obtain new methods in the following way:

We select a sequence of values $x = x_n \rightarrow \xi$, and associate with (1.2) a series-to-sequence transform

$$(1.6) \quad A_n(x_n) = A_n = \sum_{\nu=0}^n u_\nu \varphi_\nu(x_n), \quad n = 0, 1, 2, \dots;$$

its matrix is of triangular type. The summability method defined by

¹ Presented to the American Mathematical Society, September 2, 1941.

² For a comprehensive discussion of the general regular transforms cf. Szász [8], Agnew [1]; particular methods of the type (1.5) have been discussed by Perron [3]. Numbers in brackets refer to the literature listed at the end of this paper.

$$(1.7) \quad \lim_{n \rightarrow \infty} A_n = s$$

is regular (by the theorem of Silverman and Toeplitz) if and only if:

$$(1.8) \quad \lim_{n \rightarrow \infty} \varphi_\nu(x_n) = 1 \quad \text{for } \nu = 0, 1, 2, \dots,$$

$$(1.9) \quad \sum_{\nu=0}^{n-1} |\varphi_\nu(x_n) - \varphi_{\nu+1}(x_n)| + |\varphi_n(x_n)| \text{ is uniformly bounded in } n.$$

It is obvious from our regularity conditions that the method (1.7) is regular whenever (1.5) is, but the converse is not true in general.

We can apply the same generating process to a *sequence-to-function transform*:

$$(1.10) \quad \Psi(x) = \sum_{\nu=0}^{\infty} s_\nu \psi_\nu(x),$$

where $\{\psi_\nu(x)\}$ is a suitably given sequence of functions. From (1.10) we get the summability method:

$$(1.11) \quad \lim_{x \rightarrow \xi} \Psi(x) = s = \text{gen. lim. } s_n.$$

This method is regular (i.e. $\lim s_n = s$ implies (1.11)) if and only if:

$$(1.12) \quad \lim_{x \rightarrow \xi} \psi_\nu(x) = 0, \quad \text{for } \nu = 0, 1, 2, \dots,$$

$$(1.13) \quad \sum_{\nu=0}^{\infty} |\psi_\nu(x)| \text{ uniformly bounded on } R,$$

$$(1.14) \quad \sum_{\nu=0}^{\infty} \psi_\nu(x) \rightarrow 1 \quad \text{as } x \rightarrow \xi.$$

We now associate with (1.10) and with a sequence of values $x = x_n \rightarrow \xi$ a triangular-type sequence to sequence transform:

$$(1.15) \quad B_n(x_n) = B_n = \sum_{\nu=0}^n s_\nu \psi_\nu(x_n), \quad n = 0, 1, 2, \dots$$

The summability method

$$(1.16) \quad \lim B_n = s = \text{gen. lim } s_n$$

is regular if and only if:

$$(1.17) \quad \lim_{n \rightarrow \infty} \psi_\nu(x_n) = 0, \quad \text{for } \nu = 0, 1, 2, \dots,$$

$$(1.18) \quad \sum_{\nu=0}^n |\psi_\nu(x_n)| \text{ uniformly bounded in } n,$$

$$(1.19) \quad \lim_{n \rightarrow \infty} \sum_{\nu=0}^n \psi_\nu(x_n) = 1.$$

It is seen easily that regularity of the method (1.11) does not imply regularity of the method (1.15), and conversely.

Note that A_n can also be written as a sequence to sequence transform:

$$A_n = \sum_{\nu=0}^{n-1} s_\nu \{ \varphi_\nu(x_n) - \varphi_{\nu+1}(x_n) \} + s_n \varphi_n(x_n),$$

where

$$s_n = \sum_0^n u_\nu, \quad n = 0, 1, 2, \dots$$

A generalization of (1.6) is

$$A_m(x_n) = \sum_{\nu=0}^m u_\nu \varphi_\nu(x_n), \quad \text{where } m = m(n) \rightarrow \infty,$$

and the analogous sequence-to-function transform

$$F(x) = \sum_{\nu \leq m(x)} u_\nu \varphi_\nu(x).$$

These generalizations will not be discussed in the present paper.

2. We consider first the important case of Abel-summability (generalized by Stolz to complex values of x); it can be written either as a convergence-factor method:

$$\lim_{x \rightarrow 1} \sum_0^\infty u_\nu x^\nu = s,$$

or as a sequence-to-function transform:

$$\lim_{x \rightarrow 1} \sum_0^\infty s_\nu x^\nu (1-x) = s.$$

Accordingly:

$$(2.1) \quad \varphi_n(x) = x^n, \quad n = 0, 1, 2, \dots,$$

and

$$(2.2) \quad \psi_n(x) = (1-x)x^n, \quad n = 0, 1, 2, \dots$$

The method is known to be regular if and only if x approaches 1 along a path inside and non-tangential to the unit circle, i.e.

$$|x| < 1 \quad \text{and} \quad 1-x = O(1-|x|) \quad \text{as } x \rightarrow 1.$$

Writing $x = \rho e^{i\theta}$ the latter condition becomes $\theta = O(1-\rho)$.

Now the associated series to sequence transform is:

$$(2.3) \quad A_n = \sum_{\nu=0}^n u_\nu x_n^\nu = \sum_{\nu=0}^{n-1} s_\nu x_n^\nu (1 - x_n) + s_n x_n^n, \quad n = 0, 1, 2, \dots$$

The regularity conditions are (from (1.8) and (1.9)):

$$(2.4) \quad x_n \rightarrow 1 \quad \text{as } n \rightarrow \infty,$$

$$(2.5) \quad |x_n|^n + \frac{1 - |x_n|^n}{1 - |x_n|} |1 - x_n| = O(1) \quad \text{as } n \rightarrow \infty.^3$$

On the other hand, using (2.2) we get the sequence-to-sequence transform

$$(2.6) \quad B_n = \sum_{\nu=0}^n s_\nu x_n^\nu (1 - x_n),$$

and the summability method

$$\lim_{n \rightarrow \infty} B_n = s = \text{gen. lim } s_n,$$

which is regular if and only if:

$$(2.7) \quad x_n \rightarrow 1, \quad x_n^n \rightarrow 0, \quad \text{and} \quad 1 - x_n = O(1 - |x_n|) \quad \text{as } n \rightarrow \infty.$$

Although the methods $\lim A_n$ and $\lim B_n$ both originate from Abel's method in much the same way, they have quite different regularity conditions.

On writing

$$x_n = \rho_n e^{i\theta_n}, \quad -\pi < \theta_n \leq \pi,$$

the conditions (2.4) and (2.5) take the form

$$(2.8) \quad \rho_n \rightarrow 1, \quad \theta_n \rightarrow 0, \quad \rho_n^n = O(1),$$

and

$$(2.9) \quad 1 - 2\rho_n \cos \theta_n + \rho_n^2 = O\left(\left(\frac{1 - \rho_n}{1 - \rho_n^n}\right)^2\right) \quad \text{as } n \rightarrow \infty.$$

But

$$1 - 2\rho_n \cos \theta_n + \rho_n^2 = (1 - \rho_n)^2 + 4\rho_n \sin^2 \frac{1}{2}\theta_n,$$

and in view of (2.8)

$$1 - \rho_n = O\left(\frac{1 - \rho_n}{1 - \rho_n^n}\right) \quad \text{as } n \rightarrow \infty,$$

and

$$\sin \frac{1}{2}\theta_n \sim \frac{1}{2}\theta_n,$$

hence (2.9) reduces to

$$\theta_n = O\left(\frac{1 - \rho_n}{1 - \rho_n^n}\right) \quad \text{as } n \rightarrow \infty.$$

³ If $|x_n| = 1$, then $\frac{1 - |x_n|^n}{1 - |x_n|}$ is to be replaced by n .

Note that

$$\frac{1 - \rho_n}{1 - \rho_n^n} = \frac{1}{1 + \rho_n + \dots + \rho_n^{n-1}} = o(1) \quad \text{as } n \rightarrow \infty.$$

Finally, putting $\rho_n = 1 + \delta_n$; we have $\delta_n \rightarrow 0$, and $\rho_n^n = e^{n \log(1+\delta_n)}$; but

$$(2.10) \quad \log(1 + \delta) = \delta(1 + o(1)) \quad \text{as } \delta \rightarrow 0.$$

Hence $\rho_n^n = O(1)$ if and only if $n(\rho_n - 1) < k$.⁴ Summarizing we have

THEOREM 1. *Necessary and sufficient conditions for the regularity of the transform (2.3) (where $x_n = \rho_n e^{i\theta_n}$) are:*

$$(2.11) \quad \rho_n \rightarrow 1, \quad \overline{\lim} n(\rho_n - 1) < +\infty, \quad \theta_n = O\left(\frac{1 - \rho_n}{1 - \rho_n^n}\right), \quad \text{as } n \rightarrow \infty.$$

The last condition is certainly satisfied if $\theta_n = O(1/n)$; this condition is also a necessary one if $\rho_n \geq 1$ for all large n .

For the regularity of the method (2.6) the necessary and sufficient conditions are (from (2.7))

$$\rho_n \rightarrow 1, \quad \rho_n^n \rightarrow 0 \quad \text{and} \quad \theta_n = O(1 - \rho_n) \quad \text{as } n \rightarrow \infty;$$

here the second condition can be replaced by (using (2.10))

$$n(1 - \rho_n) \rightarrow +\infty \quad \text{as } n \rightarrow \infty.$$

It is easily seen that the last condition of (2.11) can be replaced by $\theta_n = O(n^{-1} + |1 - \rho_n|)$.

3. We get a summability method related to Abel's on choosing

$$\varphi_n(x) = \Re x^n = \rho^n \cos n\theta, \quad n = 0, 1, 2, \dots$$

The regularity conditions (1.3) and (1.4) now become:

$$\rho \rightarrow 1, \quad \theta \rightarrow 0 \quad \text{and}$$

$$(3.1) \quad \sum_{\nu=0}^{\infty} \rho^\nu |\cos \nu\theta - \rho \cos(\nu+1)\theta| \text{ uniformly bounded in } \rho \text{ and } \theta.$$

In particular we must have $\rho < 1$; furthermore, using the formulae

$$(3.2) \quad \cos \nu\theta - \rho \cos(\nu+1)\theta = (1 - \rho) \cos \nu\theta + 2\rho \sin \frac{1}{2}\theta \sin(\nu + \frac{1}{2})\theta,$$

and

$$\sum_0^\infty \rho^\nu (1 - \rho) |\cos \nu\theta| \leq 1,$$

(3.1) is equivalent to:

$$|\theta| \sum_0^\infty \rho^\nu |\sin(\nu + \frac{1}{2})\theta| \text{ uniformly bounded in } \rho \text{ and } \theta.$$

⁴ k, k_1, k_2, \dots denote absolute constants.

Thus a sufficient condition is:

$$\frac{\theta}{1-\rho} = O(1) \quad \text{for } \rho \rightarrow 1, \text{ or for } \theta_n^v \rightarrow 0.$$

This condition is also necessary, as is seen from

$$\begin{aligned} \sum_0^\infty \rho^v |\sin(\nu + \tfrac{1}{2})\theta| &> \sum_0^\infty \rho^v \sin^2(\nu + \tfrac{1}{2})\theta = \frac{1}{2} \left\{ \frac{1}{1-\rho} - \frac{(1-\rho)\cos\theta}{1-2\rho\cos\theta+\rho^2} \right\} \\ &= \frac{1-(1+\rho^2)\cos\theta+\rho^2}{2(1-\rho)(1-2\rho\cos\theta+\rho^2)} = \frac{(1+\rho)^2(1-\cos\theta)}{2(1-\rho)\{(1-\rho)^2+2\rho(1-\cos\theta)\}}. \end{aligned}$$

Thus the regularity condition here is the same as for Abel-summability.

The method corresponding to (1.6) is now

$$\begin{aligned} (3.3) \quad A_n(\rho_n, \theta_n) &\equiv \sum_{v=0}^n u_v \rho_n^v \cos v\theta_n \\ &= \sum_0^{n-1} s_v \rho_n^v \{\cos v\theta_n - \rho_n \cos(\nu+1)\theta_n\} + s_n \rho_n^n \cos n\theta_n; \end{aligned}$$

the regularity conditions for this method are:

$$(3.4) \quad \rho_n \rightarrow 1, \quad \theta_n \rightarrow 0 \quad \text{as } n \rightarrow \infty,$$

and

$$(3.5) \quad \rho_n^n |\cos n\theta_n| + \sum_0^{n-1} \rho_n^v |\cos v\theta_n - \rho_n \cos(\nu+1)\theta_n| = O(1) \quad \text{as } n \rightarrow \infty.$$

From $|\Re(x^v - x^{v+1})| \leq |x^v - x^{v+1}|$ it is clear that the conditions (2.11) are sufficient for the regularity of the method (3.3). We shall prove that they are also necessary. We first show that $\rho_n^n = O(1)$ is a necessary condition. If $n|\theta_n| \leq \frac{\pi}{4}$, then $\rho_n^n |\cos n\theta_n| \geq \rho_n^n \frac{1}{\sqrt{2}}$. If $n|\theta_n| > \frac{\pi}{4}$, then define $\lambda = \lambda_n \geq 1$ by

$$(2\lambda - 1)\frac{\pi}{4} < n|\theta_n| \leq (2\lambda + 1)\frac{\pi}{4},$$

and define κ as the smallest integer for which

$$\kappa|\theta_n| \geq (2\lambda - 1)\frac{\pi}{4}.$$

Thus $(\kappa - 1)|\theta_n| < (2\lambda - 1)\frac{\pi}{4} < n|\theta_n|$, $\kappa < n$; and

$$(3.6) \quad \kappa \geq (2\lambda - 1) \frac{\pi}{4|\theta_n|} = \frac{(2\lambda + 1)\pi}{4|\theta_n|} \cdot \frac{2\lambda - 1}{2\lambda + 1} \geq \frac{2\lambda - 1}{2\lambda + 1} n \geq \frac{n}{3}.$$

Furthermore

$$\begin{aligned} \sum_0^{n-1} \rho_n^r |\cos \nu \theta_n - \rho_n \cos (\nu + 1) \theta_n| \\ \geq \left| \sum_0^{n-1} (\rho_n^r \cos \nu \theta_n - \rho_n^{r+1} \cos (\nu + 1) \theta_n) \right| = |1 - \rho_n^r \cos \kappa \theta_n|, \end{aligned}$$

and

$$(2\lambda - 1) \frac{\pi}{4} \leq \kappa |\theta_n| < (2\lambda - 1) \frac{\pi}{4} + |\theta_n|,$$

hence

$$\kappa |\theta_n| = (2\lambda - 1) \frac{\pi}{4} + \epsilon_n, \quad 0 \leq \epsilon_n < |\theta_n|,$$

$$\cos \kappa |\theta_n| \geq \sin \left(\frac{\pi}{4} - |\theta_n| \right).$$

This and (3.6) proves that (3.5) implies $\rho_n^n = O(1)$. Now using (3.2) we see that (3.5) is equivalent to

$$\theta_n \sum_0^{n-1} \rho_n^r |\sin (\nu + \tfrac{1}{2}) \theta_n| = O(1) \quad \text{as } n \rightarrow \infty.$$

Hence we must have

$$(3.7) \quad \theta_n \sum_0^{n-1} \rho_n^r \sin^2 (\nu + \tfrac{1}{2}) \theta_n = O(1) \quad \text{as } n \rightarrow \infty.$$

But

$$\begin{aligned} \sum_0^{n-1} \rho^r \sin^2 (\nu + \tfrac{1}{2}) \theta &= \sum_0^{n-1} \rho^r (1 - \cos (2\nu + 1) \theta) = \frac{1 - \rho^n}{1 - \rho} - \frac{(1 - \rho) \cos \theta}{1 + \rho^2 - 2\rho \cos 2\theta} \\ &\quad + \frac{\rho^n [\cos (2n + 1) \theta - \rho \cos (2n - 1) \theta]}{1 + \rho^2 - 2\rho \cos 2\theta} \\ &= \frac{1 - \rho^n}{1 - \rho} - \frac{(1 - \rho) \cos \theta}{(1 - \rho)^2 + 4\rho \sin^2 \theta} + \frac{\rho^n [\cos (2n + 1) \theta - \cos (2n - 1) \theta]}{(1 - \rho)^2 + 4\rho \sin^2 \theta} \\ &\quad + \frac{\rho^n (1 - \rho) \cos (2n - 1) \theta}{(1 - \rho)^2 + 4\rho \sin^2 \theta}. \end{aligned}$$

Now

$$\begin{aligned} \frac{|1 - \rho_n| |\cos \theta_n|}{(1 - \rho_n)^2 + 4\rho_n \sin^2 \theta_n} &\leq \frac{|1 - \rho_n|}{4\rho_n^{\frac{1}{2}} |1 - \rho_n| |\sin \theta_n|} = O\left(\frac{1}{\theta_n}\right), \\ \frac{|\cos (2n + 1) \theta_n - \cos (2n - 1) \theta_n|}{(1 - \rho_n)^2 + 4\rho_n \sin^2 \theta_n} &\leq \frac{2 |\sin \theta_n|}{4\rho_n \sin^2 \theta_n} = O\left(\frac{1}{\theta_n}\right), \end{aligned}$$

and we find that (3.7) implies $\theta_n \frac{1 - \rho_n^n}{1 - \rho_n} = O(1)$. We have thus proved:

THEOREM 2. *Necessary and sufficient conditions for the regularity of the transform (3.3) are conditions (2.11). In the special case $\rho_n \equiv 1$ these conditions reduce to $\theta_n = O(1/n)$ as $n \rightarrow \infty$. The corresponding transform $A_n(1, \theta_n) = \sum_0^n u_\nu \cos \nu \theta_n$ was first introduced by Rogosinski [5, 6], in connection with trigonometric series.*

4. Finally let

$$\varphi_n(x) = \frac{\sin nx}{nx}, \quad n = 0, 1, 2, \dots, (\varphi_0 \equiv 1).$$

Now (1.2) with $x \rightarrow 0$ is Lebesgue's summability; it is not regular, as (1.4) is not satisfied. The associated transform (1.6) becomes

$$(4.1) \quad A_n(x_n) = \sum_0^n u_\nu \frac{\sin \nu x_n}{\nu x_n}, \quad n = 0, 1, 2, \dots$$

The necessary and sufficient conditions for regularity are (from (1.8) and (1.9) for real x_n):

$$x_n \rightarrow 0$$

and

$$\sum_0^n \left| \frac{\sin \nu x_n}{\nu x_n} - \frac{\sin (\nu+1)x_n}{(\nu+1)x_n} \right| = O(1) \quad \text{as } n \rightarrow \infty.$$

We restrict our discussion to real positive $x_n = \theta_n \downarrow 0$; we shall prove that as in Theorem 2, $n\theta_n = O(1)$ is the necessary and sufficient condition for regularity. Elementary calculus shows that $\sin \theta/\theta$ is monotonic in the successive intervals $t_{\nu-1} \leq \theta \leq t_\nu$, $\nu = 1, 2, 3, \dots$, where $t_0 = 0$ and t_1, t_2, t_3, \dots are the positive roots of the equation

$$(4.2) \quad \sin \theta = \theta \cos \theta;$$

also

$$t_\nu = \nu\pi + \alpha_\nu, \quad \text{where } 0 < \alpha_\nu < \frac{\pi}{2},$$

and substitution into (4.2) yields easily

$$\frac{\pi}{2} - \alpha_\nu < \frac{1}{2\nu},$$

in particular $\alpha_\nu \rightarrow \pi/2$. Subdividing in

$$\sum_0^n \left| \frac{\sin \nu \theta_n}{\nu \theta_n} - \frac{\sin (\nu+1)\theta_n}{(\nu+1)\theta_n} \right|$$

the summation into parts, in each of which the differences have constant sign we find now that $n\theta_n = O(1)$ is the necessary and sufficient condition for uniform boundedness. Thus we have proved:

THEOREM 3. *The transform (4.1) with $x_n = \theta_n \downarrow 0$ is regular if and only if $n\theta_n = O(1)$ as $n \rightarrow \infty$.*

5. In this section we establish a lemma for later application. We restrict ourselves to real positive sequences $\{x_n\}$. We have seen already (§2) that the class

$$(5.1) \quad x_n^n = O(1) \quad \text{as } n \rightarrow \infty,$$

is characterized by

$$(5.2) \quad \lim_{n \rightarrow \infty} n(1 - x_n) > -\infty;$$

and the smaller class

$$(5.3) \quad x_n^n = o(1) \quad \text{as } n \rightarrow \infty,$$

is characterized by

$$(5.4) \quad n(1 - x_n) \rightarrow +\infty \quad \text{as } n \rightarrow \infty.$$

We now prove the

LEMMA. *Let $0 < x_n < 1$ for all large n ; then any one of the three conditions:*

$$(5.5) \quad x_n^n \log \frac{1}{1 - x_n} = O(1),$$

$$(5.6) \quad x_n^n = O\left(\frac{1}{\log n}\right), \quad \text{as } n \rightarrow \infty,$$

$$(5.7) \quad \lim_{n \rightarrow \infty} \{n(1 - x_n) - \log \log n\} > -\infty,$$

implies the two others.

First assume (5.5); then

$$(5.8) \quad k_1 x_n^{-n} + \log(1 - x_n) > 0 \quad \text{for all large } n.$$

Let

$$g(x) = k_1 x^{-n} + \log(1 - x), \quad 0 < x < 1,$$

then

$$g'(x) = -nk_1 x^{-n-1} - \frac{1}{1-x} < 0,$$

hence $g(x) \downarrow -\infty$ as $x \uparrow 1$. From (5.8): $g(x_n) > 0$ for all large n . We introduce for a constant k

$$x_n(k) \equiv 1 - \frac{1}{n}(\log \log n - k);$$

(2.10) yields

$$(5.9) \quad (x_n(k))^n = e^{n \log x_n(k)} = \exp \{ (k - \log \log n)(1 + o(1)) \} = \frac{e^k}{\log n} (1 + o(1))$$

as $n \rightarrow \infty$. Hence $g(x_n(k)) < 0$ for $e^{-k} < 1/k_1$ and all large n . Thus

$$(5.10) \quad x_n < x_n(k) \text{ for } k > \log k_1, \text{ and for all large } n,$$

which is (5.7); moreover (5.9) and (5.10) yield (5.6). We now assume (5.6); thus

$$x_n^n < k_2 / \log n \text{ for all large } n,$$

or

$$x_n^{-n} > \log n / k_2, \text{ and } e^{n(x_n^{-n}-1)} > x_n^{-n} > \log n / k_2 \text{ for } n > k_3.$$

Hence

$$n(x_n^{-n} - 1) > \log \log n - \log k_2,$$

or

$$(5.11) \quad x_n < \frac{1}{1 + \frac{1}{n} (\log \log n - \log k_2)} < 1 - \frac{1}{n} (\log \log n - \log k_2),$$

which gives (5.7). From here we get easily

$$(5.12) \quad \log \frac{1}{1 - x_n} = \log n + O(1),$$

which together with (5.6) yields (5.5). Finally assume (5.7); then (5.11) holds, and the inequality

$$1 - x < e^{-x} \text{ for } 0 < x < 1$$

gives $x_n^n < k_2 / \log n$, which is (5.6). This and (5.12) finally give (5.5). This proves the lemma.

6. We now apply the summability methods introduced in §§2, 3, 4, to Fourier series. Let $f(\theta)$ be an integrable function of period 2π with the Fourier series:

$$(6.1) \quad f(\theta) \sim \frac{1}{2}a_0 + \sum_1^\infty (a_\nu \cos \nu\theta + b_\nu \sin \nu\theta),$$

so that

$$b_0 = 0, \quad a_\nu + ib_\nu = \frac{1}{\pi} \int_{-\pi}^{\pi} f(\theta) e^{i\nu\theta} d\theta, \quad \nu = 0, 1, 2, \dots$$

Applying (2.3) to the series (6.1) yields

$$(6.2) \quad \begin{aligned} A_n(x_n, f) &= \frac{1}{2}a_0 + \sum_{\nu=1}^n x_n^\nu (a_\nu \cos \nu\theta + b_\nu \sin \nu\theta) \\ &= \frac{1}{\pi} \int_{-\pi}^{\pi} f(\theta + t) \left(\frac{1}{2} + \sum_{\nu=1}^n x_n^\nu \cos \nu t \right) dt \\ &= \frac{1}{\pi} \int_0^\pi \{ f(\theta + t) + f(\theta - t) \} T_n(x_n, t) dt, \end{aligned}$$

where

$$T_n(x_n, t) = \frac{1}{2} + \sum_1^n x'_n \cos \nu t.$$

Let

$$H(x, \theta) = \frac{1}{2}a_0 + \sum_1^\infty x'(a_\nu \cos \nu\theta + b_\nu \sin \nu\theta),$$

$$S_n(x, \theta) = \frac{1}{2}a_0 + \sum_1^n x'(a_\nu \cos \nu\theta + b_\nu \sin \nu\theta),$$

then $A_n(x_n, f) = S_n(x_n, \theta)$.

(6.2) is an integral-transform of the function $f(\theta)$ into a sequence A_n by the kernel $T_n(x_n, t)$. Such a transform is called regular, if $A_n \rightarrow \frac{1}{2}\{f(\theta) + f(-\theta)\}$ at every point of continuity. On defining [cf. Szász [8], chapter 3]

$$K(x, t) = \frac{2}{\pi} T_n(x_n, t) \quad \text{for } n \leq x < n+1, \quad \text{and } 0 \leq t \leq \pi,$$

$$\text{and } K(x, t) = 0 \quad \text{for } x \geq 0, t > \pi,$$

furthermore

$$\chi(t) = \frac{1}{2}\{f(\theta + t) + f(\theta - t)\},$$

we get for the transform (6.2) $\int_0^\pi K(x, t)\chi(t) dt$, a step function of x . The necessary and sufficient conditions for regularity of this transform are [cf. Agnew 1].

$$(6.3) \quad \lim_{n \rightarrow \infty} \int_h^\pi T_n(x_n, t)g(t) dt = 0 \quad \text{for any integrable } g(t) \text{ and any } h > 0,$$

$$(6.4) \quad \overline{\lim}_{n \rightarrow \infty} \int_0^\pi |T_n(x_n, t)| dt < \infty,$$

$$(6.5) \quad \lim_{n \rightarrow \infty} \int_0^\pi T_n(x_n, t) dt = \frac{\pi}{2}.$$

It is of interest to consider real positive sequences x_n , for which $T_n(x_n, t)$ is non negative for all t . Evidently there exists an r_n such that $T_n(x_n, t) \geq 0$ for all t , when $0 < x_n \leq r_n$, whereas for any $\epsilon > 0$ $T_n(r_n + \epsilon, t)$ becomes < 0 for some t . I. Schur and G. Szegő [7] have proved that:

$$(6.6) \quad r_n \uparrow 1, \quad \frac{n}{\log n} (1 - r_n) \rightarrow 1 \quad \text{as } n \rightarrow \infty,$$

$$(6.7) \quad r_n > 1 - \frac{1}{n} \log 2n \quad \text{for } n \geq 1.$$

For odd n r_n is the positive root of the equation

$$(6.8) \quad 1 - r - 2r^{n+1} = 0.$$

It follows from (6.8) and (6.6) that

$$(6.9) \quad r_n^n = O(1 - r_n) = O\left(\frac{\log n}{n}\right) \text{ as } n \rightarrow \infty.$$

Thus (from §2) the transforms (2.3) and (2.6) are both regular for $x_n = r_n$. Moreover $A_n(r_n, f)$ is a regular transform; this follows from the statement:

THEOREM 4. *The function-to-sequence transform (6.2) (where $0 < x_n < 1$) is regular if and only if $x_n \rightarrow 1$ and if (5.7) holds.*

Evidently for any x_n

$$\int_0^\pi T_n(x_n, t) dt = \frac{\pi}{2},$$

hence we need only satisfy (6.3) and (6.4). We restrict ourselves to $0 < x_n < 1$, and make use of the formula [7, §1]:

$$(6.10) \quad \begin{aligned} & 2T_n(x, t) \\ &= (1 - 2x \cos t + x^2)^{-1} \{1 - x^2 + 2x^{n+2} \cos nt - 2x^{n+1} \cos (n+1)t\}. \end{aligned}$$

The expression $\frac{2}{\pi} \int_0^\pi |T_n(x_n, t)| dt = l(x_n)$ may be called the n^{th} Lebesgue constant of our summability method. As

$$\int_0^\pi \frac{1 - x^2}{1 - 2x \cos t + x^2} dt = \pi,$$

it is clear that (6.4) is satisfied if and only if

$$(6.11) \quad x_n^n \int_0^\pi \frac{|x_n \cos nt - \cos (n+1)t|}{1 - 2x_n \cos t + x_n^2} dt = O(1) \quad \text{as } n \rightarrow \infty.$$

But from the identity

$$x \cos nt - \cos (n+1)t = x\{\cos nt - \cos (n+1)t\} - (1-x) \cos (n+1)t$$

it follows that (6.11) holds if and only if

$$(6.12) \quad x_n^n \int_0^\pi |\cos nt - \cos (n+1)t| (1 - 2x_n \cos t + x_n^2)^{-1} dt = O(1).$$

We now prove that the necessary and sufficient condition for (6.12) is (5.5). To show that (5.5) implies (6.12) write

$$\begin{aligned} & \int_0^\pi |\cos nt - \cos (n+1)t| (1 - 2x_n \cos t + x_n^2)^{-1} dt \\ &= 2 \int_0^\pi \sin \frac{t}{2} \frac{|\sin (n + \frac{1}{2})t| dt}{(1 - x_n)^2 + 4x_n \sin^2 \frac{t}{2}} < \int_0^\pi \frac{t dt}{(1 - x_n)^2 + \frac{4}{\pi^2} x_n t^2} \\ &< \int_0^{1-x_n} \frac{t dt}{(1 - x_n)^2} + \frac{2}{x_n} \int_{1-x_n}^\pi t^{-1} dt = \frac{1}{2} + \frac{2}{x_n} \log \frac{\pi}{1 - x_n}, \end{aligned}$$

for $0 < x_n < 1$; this proves the first part of our statement. The converse is seen from the estimate:

$$\int_0^\pi \sin \frac{t}{2} \frac{|\sin(n + \frac{1}{2})t| dt}{(1 - x_n)^2 + 4x_n \sin^2 \frac{t}{2}} > \frac{1}{2\pi} \int_{1-x_n}^\pi \frac{t |\sin(n + \frac{1}{2})t| dt}{t^2} > k \log \frac{1}{1 - x_n}.$$

We finally show that (6.3) is satisfied whenever $x_n \rightarrow 1$ (assuming that (5.5) holds). We have

$$x_n^n \int_h^\pi |x_n \cos nt - \cos(n+1)t| |g(t)| (1 - 2x_n \cos t + x_n^2)^{-1} dt < \frac{x_n^n}{1 - \cos h} \int_0^\pi |g(t)| dt = o(1)$$

as $n \rightarrow \infty$ (from (5.6)). Hence, using (6.10), (6.3) holds if and only if

$$(1 - x_n^2) \int_h^\pi (1 - 2x_n \cos t + x_n^2)^{-1} g(t) dt \rightarrow 0 \quad \text{as } n \rightarrow \infty,$$

and this is equivalent to $x_n \rightarrow 1$. From the lemma of §5 we finally get Theorem 4.

We now apply the summability method (2.6) to Fourier series. Introducing the notation

$$s_0 = \frac{1}{2}a_0, \quad s_n(f, \theta) = s_n = \frac{1}{2}a_0 + \sum_1^n (a_\nu \cos \nu\theta + b_\nu \sin \nu\theta), \quad n \geq 1,$$

and using the formula

$$s_n = \frac{1}{2\pi} \int_0^\pi \{f(\theta + t) + f(\theta - t)\} \frac{\sin(n + \frac{1}{2})t}{\sin \frac{1}{2}t} dt,$$

we get from (2.6) the function-to-sequence transform:

$$(6.13) \quad B_n(x_n, f) \equiv \frac{1 - x_n}{2\pi} \int_0^\pi \{f(\theta + t) + f(\theta - t)\} U_n(x_n, t) dt,$$

where

$$U_n(x_n, t) = \frac{1}{\sin \frac{1}{2}t} \sum_1^n x_n^\nu \sin(\nu + \frac{1}{2})t.$$

For the regularity of the transform (6.13) we have the necessary and sufficient conditions:

$$(6.14) \quad \lim_{n \rightarrow \infty} (1 - x_n) \int_h^\pi U_n(x_n, t) g(t) dt = 0 \quad \text{for any } g(t) \in L,$$

and any given $h > 0$,

$$(6.15) \quad \overline{\lim}_{n \rightarrow \infty} \frac{1 - x_n}{\pi} \int_0^\pi |U_n(x_n, t)| dt < \infty,$$

$$(6.16) \quad \lim_{n \rightarrow \infty} \frac{1 - x_n}{\pi} \int_0^\pi U_n(x_n, t) dt = 1.$$

For $f(t) \equiv 1$ formula (6.13) yields

$$1 - x_n^{n+1} = \frac{1 - x_n}{\pi} \int_0^\pi U_n(x_n, t) dt,$$

and condition (6.16) becomes $x_n^n \rightarrow 0$ as $n \rightarrow \infty$; assume again $0 < x_n < 1$, then we must have $n(1 - x_n) \rightarrow +\infty$. But Fejér [2, p. 20] remarked that for any $c_0 > c_1 > \dots > c_n > 0$, $\sum_0^n c_\nu \sin(\nu + \frac{1}{2})t > 0$ for $0 < t < 2\pi$; hence the kernel of the transform (6.13) is positive, and the condition (6.15) is satisfied. We finally discuss condition (6.14). From the harmonic development

$$\sum_0^\infty x^\nu \sin(\nu + \frac{1}{2})t = \frac{(1+x) \sin \frac{1}{2}t}{1 - 2x \cos t + x^2}, \quad 0 < x < 1,$$

we find easily the formula

$$\begin{aligned} U_n(x, t) \sin \frac{1}{2}t (1 - 2x \cos t + x^2) \\ = (1+x) \sin \frac{1}{2}t + x^{n+2} \sin(n + \frac{1}{2})t - x^{n+1} \sin(n + \frac{3}{2})t. \end{aligned}$$

Using this formula and the condition $x_n^n \rightarrow 0$, it follows immediately that (6.14) holds if and only if

$$(1 - x_n) \int_h^\pi (1 - 2x_n \cos t + x_n^2)^{-1} g(t) dt \rightarrow 0,$$

i.e. if $x_n \rightarrow 1$. Thus we have

THEOREM 5. *The necessary and sufficient conditions for the regularity of the transform (6.13) (where $0 < x_n < 1$) are:*

$$x_n \rightarrow 1 \quad \text{and} \quad n(1 - x_n) \rightarrow +\infty, \quad \text{as } n \rightarrow \infty.$$

7. We give here a Tauberian type theorem for the method $A_n(x_n)$ of §2.

THEOREM 6. *Suppose x_n real or complex, and*

$$(7.1) \quad \overline{\lim}_{n \rightarrow \infty} |1 - x_n| \frac{|x_n|^{-n} - 1}{1 - |x_n|} = \lambda < 1,$$

then $A_n(x_n) \rightarrow s$ implies $s_n \rightarrow s$.

Note that (7.1) implies $\overline{\lim}_{n \rightarrow \infty} ||x_n|^{-n} - 1| \leq \lambda$, and in particular $x_n \rightarrow 1$, and

$$\frac{1}{1+\lambda} \leq \underline{\lim}_{n \rightarrow \infty} |x_n|^n \leq \overline{\lim}_{n \rightarrow \infty} |x_n|^n \leq \frac{1}{1-\lambda}.$$

The theorem can be deduced from recent results of R. Rado [4]; we give a direct proof, following his device.

Suppose first that $s = 0$ and $s_n = O(1)$; then $0 \leq \overline{\lim}_{n \rightarrow \infty} |s_n| = \delta < \infty$; we shall prove that the assumption $\delta > 0$ leads to a contradiction. To a given $\epsilon > 0$ choose $n = n(\epsilon)$ so that $|s_\nu| < \delta + \epsilon$ for $\nu > n$; then choose $m > n$ so that $|s_m| > \delta - \epsilon$. Now, using (2.3)

$$\begin{aligned} \delta - \epsilon < |s_m| &= \left| A_m x_m^{-m} - \sum_0^{m-1} s_r x_m^{r-m} (1-x_m) \right| \\ &< |A_m x_m^{-m}| + \left| \sum_0^n s_r x_m^{r-m} (1-x_m) \right| + (\delta + \epsilon) |1-x_m| \frac{|x_m|^{-m} - 1}{1 - |x_m|}; \end{aligned}$$

thus, if m is large enough

$$\delta - \epsilon < \epsilon + \epsilon + (\delta + \epsilon)(\lambda + \epsilon) < \delta\lambda + \epsilon(3 + \delta + \epsilon),$$

where $\lambda < 1$. This is a contradiction for ϵ small enough.

We next assume $s = 0$ and $\lim |s_n| = \infty$; choose $0 < \epsilon$ small and l large; denote by $m = m(l)$ the least n for which $|s_n| > l$. Then (2.3) yields

$$l < |s_m| \leq |A_m x_m^{-m}| + l |1-x_m| \frac{|x_m|^{-m} - 1}{1 - |x_m|} < \epsilon + l(\lambda + \epsilon),$$

which again is impossible for small ϵ . Thus the theorem is proved for $s = 0$. Finally for $s \neq 0$ we apply our result to the sequence $s_n - s$, and its transform

$$\sum_{r=0}^{n-1} (s_r - s) x_n^r (1-x_n) + (s_n - s)x_n^n = A_n - s \rightarrow 0,$$

which yields $s_n \rightarrow s$; our theorem is proved.

FINAL NOTE. The argument of §6 can be used to prove that the transforms $A_n(x_n, f)$, $B_n(x_n, f)$ under the regularity conditions of Theorems 4 and 5 converge to $f(\theta)$ if only

$$\int_0^t |f(\theta + u) + f(\theta - u) - 2f(\theta)| du = o(t) \quad \text{as } t \rightarrow 0,$$

thus for almost all θ .

UNIVERSITY OF CINCINNATI

LITERATURE

1. R. P. AGNEW, *Properties of generalized definitions of limit*, Bull. Am. Math. Soc., vol. 45, 1939, pp. 690-730.
2. L. FEJÉR, *Trigonometrische Reihen und Potenzreihen mit mehrfach monotoner Koeffizientenfolge*, Trans. Am. Math. Soc., vol. 39, 1936, pp. 18-59.
3. O. PERRON, *Beitrag zur Theorie der divergenten Reihen*, Math. Zeitschrift, vol. 6, 1920, pp. 286-310.
4. R. RADO, *Some elementary Tauberian Theorems* (1), Quarterly Jour. of Math., Oxford Series, vol. 9, 1938, pp. 274-282.
5. W. ROGOSINSKI, *Über die Abschnitte trigonometrischer Reihen*, Math. Annalen, vol. 95, 1926, pp. 110-134.
6. W. ROGOSINSKI, *Abschnittsverhalten bei trigonometrischen und insbesondere Fourierschen Reihen*, Math. Zeitschrift, vol. 41, 1936, pp. 75-136.
7. I. SCHUR UND G. SZEGÖ, *Über die Abschnitte einer im Einheitskreise beschränkten Potenzreihe*, Sitzungsberichte der preussischen Akademie, 1923, pp. 545-560.
8. O. SZÁSZ, *Lectures on summability methods*, University of Cincinnati, 1936-1937.

GENERALIZED SURFACES IN THE CALCULUS OF VARIATIONS

By L. C. YOUNG

(Received September 30, 1941)

I. GENERALIZED LIPSCHITZIAN SURFACES

1. Introduction

This note deals with an extension of the idea of surface, which is similar to the writer's generalization¹ of the notion of curve in the Calculus of Variations. The primary object of such generalizations is to ensure, as far as possible, that every problem of the Calculus of Variations has at least one solution. Their need has been apparent ever since the illuminating remark made by Hilbert:² *Eine jede Aufgabe der Variationsrechnung besitzt eine Lösung sobald hinsichtlich der Natur der Grenzbedingungen geeignete einschränkende Annahmen erfüllt sind, und, nötigenfalls, der Begriff der Lösung eine sinngemässe Erweiterung erfährt.*

In the existence theorems for variational problems concerning an unknown curve, the customary restriction to regular problems can now be removed by using *generalized curves*, and applications of these notions have lately been made by McShane.³

Problems of minima of variational problems concerning surfaces are *essentially* far more complicated than those concerning curves. In this note, however, the additional complication appears only in the proof of certain theorems, not in their statement.

The definition of *generalized surface* given here, and the proof of the existence of a (generalized) solution, are framed for a problem of minimum originally stated in a class Σ of ordinary surfaces

$$z = z(x, y)$$

with a given boundary. In the existence theorem, the principal further hypothesis is the restriction to a Lipschitzian class of surfaces, by which we mean that the function $z(x, y)$ satisfies the condition

$$|z(x', y') - z(x'', y'')| \leq K\rho,$$

where K is a constant depending only on the class Σ and where ρ is the distance of the arbitrary points (x', y') and (x'', y'') . The restriction expressed by this Lipschitz condition is rendered necessary by the ordinary (as opposed to the

¹ Young [12], [13], [14].

² Hilbert [2] p. 184; cf. also Lebesgue [3] p. 372.

³ McShane [4], [5], [6].

parametric) formulation of the problem, not by the fact that we have surfaces rather than curves.

It is perhaps superfluous to add that the results of the present note merely constitute a beginning. The Lipschitz condition, although needed in the general type of problem considered here, is inconvenient in the applications to classical problems such as that of Dirichlet or Plateau, where a Lipschitzian boundary function may correspond to one or more singularities of the solution. And in any case, the usefulness of existence theorems which are based on a generalization, devised for the purpose, of the notion of surface, is naturally dependent on a new extension of the necessary conditions for attained minima such as that studied by the writer and by McShane in the case of curves. These questions we reserve for a later date.

2. An elementary unattained minimum

We shall follow the example set by McShane in his treatment of generalized curves⁴ and begin with a "heuristic discussion." For the purpose of this discussion, let us compare our variational problem with an elementary minimum problem which consists in finding the least value of the function

$$f(t) = t^4 - 8t,$$

the latter being supposed defined in the first instance for *rational* t only. Let us analyze briefly this elementary problem as treated in the Differential Calculus, the term "least value" being first reinterpreted in the obvious way.

We first define *convergence* of a sequence of rational numbers, so as to make every such simple function $f(t)$ *continuous*. More precisely, whenever $\{t_n\}$ is a convergent sequence of rational numbers so is $\{f(t_n)\}$. Such a definition is furnished by the classical convergence criterion of Cauchy; indeed the classical idea of convergence may fairly be said to have been devised primarily for this very purpose. Next we define irrational numbers as limits of rational ones, and we extend to them the definition of the function $f(t)$ by continuity. In so doing we need not refer to properties special to numbers, except those implied in the definition of convergent sequence.⁵ Finally we apply the methods of elementary Analysis to show that in its extended range the given function attains its least value, and to determine the latter in the usual way.

Can we proceed similarly in the Calculus of Variations? The above comparison strongly suggests that we can, and this without restricting ourselves in

⁴ McShane [4] p. 514.

⁵ The writer has in mind a definition "by abstraction" according to the terminology of Weyl [10] p. 7; and this definition is perhaps free from the criticism which Weyl himself [11] has directed against irrationals. Two sequences have the property of *limiting equality* if they are subsequences of a same convergent sequence; a system of sequences between any two of which the above property holds defines a real number, which is termed *rational* when the system contains a sequence of repetitions, otherwise *irrational*. With this definition, only the property of continuity actually stated is needed to extend $f(t)$ to irrational values of t .

any way to regular problems as past writers have done, if only we can produce a suitable definition of convergence. Now it is quite true that the definition of convergence normally applied to surfaces S in Analysis, renders the functions $F(S)$ which occur in variational problems, discontinuous; this is so even in regular problems, when the functions are merely semi-continuous. But, as remarked by McShane, we need not look very far for a more suitable definition: it is sufficient to term the sequence of surfaces $\{S_n\}$ convergent if and only if, for every $F(S)$ of the Calculus of Variations, the sequence of numbers $\{F(S_n)\}$ converges. The functions $F(S)$ are then, automatically, continuous.

3. Connection with linear operations

In classical problems of the Calculus of Variations, the functions $F(S)$ have the form

$$L_S(f) = \int \int_S f(x, y, z, p, q) dx dy = \int \int_A \tilde{f}(x, y) dx dy,$$

where the integrand f is a continuous function of five real variables and \tilde{f} is its value at the point x, y of the domain⁶ of definition A when we substitute for the three variables z, p, q the function $z(x, y)$ and its gradient $p(x, y), q(x, y)$. We shall suppose for simplicity that the domain A is *convex*. Actually only minor verbal changes are needed to make the discussion applicable to domains of the most general kind.^{6a}

Just as in the writer's work on rectifiable curves, it is possible to treat problems of minimum in which a number of such double integrals occur simultaneously, by allowing the functions f and F to assume values in a vector-space instead of only real values; we shall content ourselves with the case of functions taking real values, and leave the reader to make the modifications appropriate to vector-functions which are suggested by the writer's treatment for curves.⁷

With the above form for $F(S)$, the definition of convergence of surfaces suggested in the preceding paragraph leads us to study linear operations $L(f)$ which are defined for every continuous function f . The definition itself, for the sequence of surfaces $\{S_n\}$, coincides with what is known as *weak convergence* of the operations $L_{S_n}(f)$, that is to say the convergence for each f of the sequence of numbers defined by these operations on f .

We shall suppose the functions f defined in the whole space of the five variables x, y, z, p, q as we clearly may; but we shall see in the next paragraph that no genuine linear operations, and no genuine weakly convergent sequences of linear operations, can exist, which are not wholly independent of the values taken by the functions f outside some fixed sphere $x^2 + y^2 + z^2 + p^2 + q^2 \leq k^2$. This

⁶ By a *domain* we mean a closed set of positive plane measure whose boundary is of plane measure zero.

^{6a} The principal of these changes consists in interpreting *distance* of two points to mean minimum length of a path in A joining them.

⁷ Young [13].

requires on the one hand that the domain A must be interior to a fixed circle of the x, y plane, and on the other hand that p, q be bounded, i.e. that the class of surfaces considered be *Lipschitzian*.

4. Linear operations in a certain type of space

Among the spaces which have become of interest in recent years, the best known are the Banach spaces, of which a particular case is the space whose elements are the continuous functions defined in a closed n -dimensional sphere. The space which is relevant in this note is one whose elements are the continuous functions defined in the *whole* of a five-dimensional space; in this case, any element f can be regarded as specified by means of a sequence $\{f_n\}$, where f_n denotes the continuous function defined in a sphere of radius n and fixed center, which coincides in its sphere with the corresponding values of f .

We formulate the results of this paragraph rather more generally, in a space whose elements f are specified by sequences $\{f_n\}$, where f_n is a point which lies in a Banach space, possibly variable with n . Writing $|f_n|$ for the norm, or magnitude, of f_n in its space, we denote by $Q_n(f)$ the greatest of the first n numbers $|f_k|$; in the special case which mainly concerns us $Q_n(f)$ is the maximum absolute value of the function in the sphere of radius n . We write further

$$(4.1) \quad Q(f) = \sum 2^{-n} \frac{Q_n(f)}{1 + Q_n(f)},$$

and we define the distance of two elements f, g to be $Q(f - g)$. In addition to verifying the usual laws of a distance, we observe at once from the definition of $Q(f)$ and from the fact that $Q_n(f)$ increases with n , that

$$(4.2) \quad \text{If } Q_N(f) \leq 2^{-N} \text{ for some } N, \text{ then } Q(f) \leq 2^{-(N-1)}.$$

This being so, let us consider the linear functions $L(f)$ defined in our space of elements f . We shall suppose the space *complete* when the distance is defined as above, a condition that is satisfied by the space of the continuous functions of five unlimited real variables. We shall further restrict each $L(f)$ to be measurable B , and therefore continuous⁸ in f . Moreover, the absolute value $|L(f)|$ being continuous also, an important result⁹ enables us to assert that, for a sequence $\{L_m(f)\}$ of such functions, which is bounded for each f of a set of the second category, there exists a sphere of elements f in which the sequence is *uniformly* bounded.

Writing $f = f_0 + g$ where f_0 is the center of this sphere, we see that the sequence $\{L_m(g)\} = \{L_m(f) - L_m(f_0)\}$ is uniformly bounded in the appropriate sphere of elements g with the origin as center, and *a fortiori* in the set $[Q_N(g) \leq 2^{-N}]$, which by (4.2) is a subset of this sphere if N is large. By homogeneity, it follows that there is a constant K such that for every f , and for all m ,

⁸ Banach [1] p. 23 theorem 4.

⁹ Banach [1] p. 19 theorem 11.

$$|L_m(f)| \leq K \cdot Q_N(f).$$

We have thus proved the following theorem.

(4.3) *If the sequence of the values of the linear functions $L_m(f)$ is bounded for each f of a set of the second category, then there exist constants K and N independent of f and m , such that*

$$(4.4) \quad |L_m(f)| \leq K \cdot Q_N(f).$$

As an immediate corollary, obtained by taking $\{L_m(f)\}$ to consist of repetitions of the linear function $L(f)$ defined for all f , we get

(4.5) *A linear function $L(f)$ defined for all the sequences $f = \{f_n\}$ depends only on a finite number of the f_n .*

In fact, if the points f and g denote two such sequences having the same N first terms, (4.4) applied to the point $f - g$, for which we have $Q_N(f - g) = 0$, gives $L_m(f - g) = 0$ where $L_m(f)$ is now a repetition of $L(f)$, and so $L(f) = L(g)$.

A set of linear functions $L(f)$ will be termed *compact* if every sequence of linear functions belonging to the set contains a subsequence which converges weakly. The following result is again a corollary of theorem (4.3).

(4.6) *The linear functions $L(f)$ of a compact set depend only on a bounded number of the f_n .*

For otherwise, we could form a sequence $L_n(f)$ of functions of the set such that (4.4) were false when $m = N = n$. This sequence would have to be unbounded at some f and so could not be compact, contrary to the hypothesis.

5. Generalized surfaces

We have already associated the notion of convergence of surfaces with the expressions $F(S)$ rather than with a geometrical representation of the surfaces concerned. Since the notion of generalized surface is to be attached to that of convergence, it is the process of calculation of the relevant $F(S)$ that we shall really be generalizing, rather than the geometrical representation.

Now this process of calculation is mainly that of passing from an arbitrary function of five variables $f(x, y, z, p, q)$ which we supposed continuous, to a function of two variables $\bar{f}(x, y)$ which we rendered measurable in A , by substituting for the three variables z, p, q the corresponding functions of x, y on S . From the function \bar{f} we then passed to $F(S)$ by integration. In generalizing this process, we place, for reasons sufficiently explained elsewhere,¹⁰ the substitution of $z(x, y)$ for z on a wholly different footing to that of its gradient for p, q . We still substitute for z a function $z(x, y)$ of which we say that it defines the track of the generalized surface S^* , but instead of substituting similarly for p, q we form a certain average of the values assumed by f at the various p, q .

By an *average*, or more precisely a *linear average* or *linear mean*, of a continuous function $g(p, q)$ of the variables p, q , we designate a functional $M(g)$,

¹⁰ Young [13] p. 231.

linear in g , non-negative for non-negative g , and which satisfies the condition $M(g) = 1$ when g is the constant unity.

We term *measurable* in x, y , an average which varies according to the values of the additional parameters x, y , if for every continuous g which depends on p, q only, the number $M(g)$ obtained by forming the average at x, y defines a measurable function of x, y . As shown by McShane,¹¹ this number then still defines a measurable function when g is a continuous function of all four variables x, y, p, q .

This being so, we are now in a position to state our definitions. A *generalized surface* is the system defined by the following pair of elements: (i) a *real function* $z(x, y)$, absolutely continuous in the sense of Tonelli;¹² (ii) an *average* M defined for points x, y of A and measurable in x, y , which operates on each continuous function of p, q , and which satisfies for almost all x, y of A the condition that for the two functions $g(p, q) = p$ and $g(p, q) = q$ the averages $M(p), M(q)$ are the components of the gradient of $z(x, y)$. It should be added that we identify two generalized surfaces when they have identical tracks $z(x, y)$ and for almost all x, y identical averages M . In the sequel, we shall often assume, without saying so explicitly, that the average has been suitably modified at the points x, y of a set of plane measure zero.

The generalized surface S^* will be termed *Lipschitzian with the constant K* , if at each point x, y the average $M(g)$ of an arbitrary function $g(p, q)$ is wholly independent of the values taken by $g(p, q)$ outside the circle of radius K whose center is the origin of the p, q plane.

With the definition of a generalized surface S^* , we give also that of *integral over S^* of the continuous function $f(x, y, z, p, q)$* , i.e. the meaning to be attached to the symbols

$$F(S^*) = L_{S^*}(f) = \int_A \tilde{f}(x, y) dx dy.$$

We do this by stipulating that $\tilde{f}(x, y) = M(g)$ where $g = g(p, q)$ is the value at the point x, y of the function $g(x, y, p, q)$ obtained by substituting $z(x, y)$ for z in $f(x, y, z, p, q)$. The fact that M is measurable in x, y ensures that the function \tilde{f} is so too.

Let us observe that an ordinary surface S and the corresponding function $F(S)$ are special cases of the above definitions: the average M then consists for almost all x, y in substituting in the function concerned the gradient of $z(x, y)$ for p, q .

6. Convergence of generalized surfaces

We now extend to the case of generalized surfaces the definition of convergence proposed in §2, and at the same time avoid conflict with the definition which is usual in Analysis. We shall consider only *Lipschitzian* generalized sur-

¹¹ McShane [4] p. 517 Lemma 3.1; the proof is given in the case of a single parameter t , but extends without difficulty to two parameters x and y

¹² Saks [8] p. 169.

faces, this restriction being necessary in view of (4.7), in order that the repetitions of a given generalized surface constitute a sequence which converges according to our definition. We say that the sequence of generalized surfaces $\{S_n^*\}$ converges if the sequence $\{L_{S_n^*}(f)\}$ converges weakly, i.e. converges for every continuous function $f(x, y, z, p, q)$.

Just as we distinguish between a generalized surface S^* and its track, we distinguish also between convergence of generalized surfaces and *convergence of their tracks*, the latter notion being defined to be uniform convergence of the corresponding functions $z(x, y)$, as in the case of the usual definition of convergence of surfaces, which we thus identify with, and shall refer to in the sequel as, convergence of the tracks of surfaces.

Let us observe further that our definition of convergent sequence of generalized surfaces contains no reference to a limit. We shall see in §7, that the space of Lipschitzian generalized surfaces is complete, so that this kind of convergence may eventually be identified with the more familiar notion of convergence to a limit. In the meantime let us state explicitly that the term *compact set* denotes a set in which all infinite subsets contain corresponding convergent sequences, and that we do not restrict these sequences to possess limits at the moment. In the writer's opinion, this interpretation of the notion of compactness is in any case preferable to the one usually adopted, since this notion then becomes an intrinsic property of a set instead of a property which depends partly on a set and partly on the space.

This being so, we have the following theorem of compactness.

(6.1) *A system of generalized surfaces is compact if and only if their tracks lie in a fixed sphere and their Lipschitz constants are bounded.*

The pair of conditions stated is clearly equivalent to the single condition which restricts the corresponding linear operations $L_{S^*}(f)$ to be wholly independent of the values of the continuous functions f outside a fixed sphere of the x, y, z, p, q space. The necessity of these conditions therefore follows from (4.6). To prove their sufficiency, we suppose them satisfied, so that the linear operations $L_{S^*}(f)$ may be regarded as defined in the *Banach space* consisting of all functions of x, y, z, p, q which are continuous in a fixed sphere; the conclusion then reduces to a corollary of a well known theorem.¹³

We now consider some consequences of the notion of convergence. If S^* is any generalized surface, and we write as usual

$$L_{S^*}(f) = \iint_A f(x, y) dx dy,$$

for the corresponding linear operation, we define more generally, for any interval Δ of the x, y plane, the *linear operation additive in Δ*

$$L_{S^*}(f; \Delta) = \iint_{A \cdot \Delta} f(x, y) dx dy.$$

¹³ Banach [1] p. 123 theorem 3.

We remark that for any fixed function f the operations corresponding to a compact system of generalized surfaces are *equi-continuous* in Δ . In fact, denoting by $K(f)$ the maximum modulus of the function f in a fixed sphere of the x, y, z, p, q space outside which the values of f are irrelevant by the preceding theorem, we clearly have

$$(6.2) \quad |L_{S^*}(f; \Delta) - L_{S^*}(f; \Delta')| \leq K(f) \cdot |\Delta - \Delta'|.$$

We shall now establish the following result.

(6.3) *If the generalized surfaces S_n^* converge, then the operations $L_{S_n^*}(f; \Delta)$ converge for each f and uniformly in Δ .*

It is sufficient by equi-continuity, to establish the convergence for each f and each Δ . We enclose the interval Δ in a concentric similar interval Δ' of slightly larger area, and we denote by g a continuous function of x, y, z, p, q whose absolute value is majorized by that of f , such that for all z, p, q we have $g = f$ when x, y lies in Δ , and $g = 0$ when x, y lies outside Δ' . The existence of a function g with these properties is clear.

This being so, it follows from (6.2) that the expressions

$$L_{S_n^*}(f; \Delta) = L_{S_n^*}(g; \Delta)$$

differ from the corresponding terms of the *convergent* sequence formed by the expressions

$$L_{S_n^*}(g) = L_{S_n^*}(g; \Delta'),$$

at most by the arbitrarily small amount

$$K(g) \cdot |\Delta' - \Delta| \leq K(f) \cdot |\Delta' - \Delta|,$$

and hence that they too must converge by Cauchy's convergence test. This completes the proof.

(6.4) *If the generalized surfaces S_n^* converge, so do their tracks.*

To see this, we observe in the first place that the functions z_n defining these tracks have by (6.1) uniformly bounded gradient, since their gradients at any point are averages of bounded sets of values of p, q . These functions are therefore equi-continuous, so that to show that they converge uniformly we need only verify that they converge at an everywhere dense set or that they converge in some average sense in the x, y plane. The latter is the case by (6.3) applied to the function $f(x, y, z, p, q) = z$, which states that the double integrals over Δ of the functions $z_n(x, y)$ converge uniformly.

7. The completeness theorem

We say that the sequence $\{S_n^*\}$ has the *limit* S^* if the operations $L_{S_n^*}(f)$ have the limit $L_{S^*}(f)$ for each f . Evidently the existence of the limit implies the convergence of the sequence; we shall establish the converse.

(7.1) *A convergent sequence of generalized surfaces has a limit.*

We denote by $z(x, y)$ the limit of the corresponding tracks, and by $L(f; \Delta)$

that of the corresponding operations additive in Δ , the limits existing by (6.4) and (6.3). We denote further by $Q(f; \Delta)$ the maximum modulus of the function f when $z = z(x, y)$, when x, y lies in $A \cdot \Delta$, and when p, q is bounded in magnitude by the upper bound of the Lipschitz constants of our sequence, finite by (6.1).

Clearly the operation $L(f; \Delta)$ is linear in f , additive in Δ , and we have

$$(7.2) \quad |L(f; \Delta)| \leq |\Delta| \cdot Q(f; \Delta),$$

as the limit of similar relations.

It follows from (7.2) that $L(f; \Delta)$ is the double integral over Δ of any network derivative $M(f; x, y)$, and that the latter exists almost everywhere, as the limit, for a suitable sequence of intervals Δ tending to the point x, y , of the expression

$$(7.3) \quad L(f; \Delta)/|\Delta|$$

when the function f is kept fixed.

For almost every x, y , this limit exists simultaneously for all functions f of a given enumerable everywhere dense set of functions. Since the functionals (7.3) are, by (7.2), equi-continuous in f for the various Δ , the derivative in question also exists simultaneously for all f at any such point x, y ; and it thus represents, for almost any fixed x, y , a linear operation in f which, as we deduce at once from (7.2), is majorized by the maximum modulus, for the same system of p, q as before, of the function f when x, y and $z = z(x, y)$ are fixed. Since it has this majorant, the operation $M(f; x, y)$ depends on the values of f as function of p, q only, the other variables being kept fixed.

This being so, let us write simply $M(f)$ for the operation just defined at almost every x, y , and let us complete its definition for the remaining points x, y by choosing it for instance to mean the value of f for $p = q = 0$.

We verify at once that $M(f)$, at the point x, y and $z = z(x, y)$, is a non-negative operation when f is non-negative for all p, q at this point, and that for the function $f = 1$ we have $M(f) = 1$. Hence $M(f)$ is an average. Moreover we have already remarked that its double integral over any Δ coincides with $L(f; \Delta)$. In order to identify the latter with the operation $L_{S^*}(f; \Delta)$, and thus to complete the proof of (7.1), it is now sufficient to verify that a generalized surface S^* is defined by the track $z(x, y)$ and the average $M(f)$. This will be the case if the averages of the functions $f = p$ and $f = q$ are the components of the gradient of $z(x, y)$ at almost all points x, y of the domain of definition A .

To establish this last result, it is sufficient to show that for these two functions f and for an arbitrary interval Δ contained in the domain A , the expression $L(f; \Delta)$ coincides with the double integral over Δ of the corresponding component of the gradient. Since the corresponding relations are true for the sequence of which L is the limit, it only remains to verify that the double integral over Δ of the gradients of the tracks of the generalized surfaces of our sequence tend to the double integral of the gradient of $z(x, y)$. This is evidently the case in view of the uniform convergence of these tracks.

8. Closure of the class of ordinary surfaces

From the theorem of completeness just established, it follows in particular that the limit of any convergent sequence of *ordinary* surfaces necessarily exists and represents a generalized surface whose Lipschitz constant is majorized by the upper bound of those occurring in the sequence. We now consider the converse question, whether *every* Lipschitzian generalized surface is expressible as the limit of a suitable sequence of ordinary surfaces. Our next theorem shows that the answer is in the affirmative, even with additional restrictions on this sequence.

(8.1) *The class of the generalized surfaces whose tracks have a fixed boundary and whose Lipschitz constants do not exceed N , is identical with the closure of the subclass consisting of ordinary surfaces.*

It is sufficient to prove that, given any generalized surface S^* with a Lipschitz constant not exceeding N , there exists a sequence of ordinary surfaces S_n with Lipschitz constants not exceeding N and boundary coinciding with that of the track of S^* , such that, for all continuous functions f of the variables x, y, z, p, q ,

$$(8.2) \quad L_{S^*}(f) = \lim_n L_{S_n}(f).$$

Let us say that a generalized surface S^{**} is an ϵ -approximation to the generalized surface S^* , if there exist a positive function η of ϵ only, which tends to 0 with ϵ , such that

$$(8.3) \quad |L_{S^*}(f) - L_{S^{**}}(f)| \leq K\eta \cdot Q(f) + K \cdot \omega(f; \eta),$$

where K is some fixed constant, and where $Q(f)$ and $\omega(f; \eta)$ denote, for the function f , the maxima, in some fixed sphere of x, y, z, p, q , of the modulus and of the oscillation between two points distant less than η .

In order to establish (8.2), and so (8.1), it is evidently enough to show that, for any positive ϵ , there is an ϵ -approximation to the given generalized surface S^* provided by an ordinary surface subject to the conditions stated for S_n . We shall divide the proof into several stages which occupy §§9–13.

9. A lemma due to McShane¹⁴

In the sequel we frequently have to enlarge and fit together disconnected portions of surfaces, without unduly increasing their Lipschitz constants. In so doing, we make use of the following lemma.

(9.1) *Suppose that a function $z(x, y)$, originally given in the plane set E , satisfies, for all pairs of points of this set, the Lipschitz condition*

$$(9.2) \quad |z(x', y') - z(x'', y'')| \leq K \cdot \rho,$$

where ρ denotes the distance of the pair of points, and where K is a fixed constant. Then there exists a function $z(x, y)$, defined in the whole plane and coinciding with

¹⁴ McShane 71 p. 838, Theorem I.

the given function in the set E , such that (9.2) continues to hold with this same value of K for all pairs of points of the plane.

For a proof of this lemma, we refer the reader to McShane.¹⁴ We shall deduce a slight refinement of some importance to us. We remark that the continuation whose existence is asserted is by no means unique, although there are certain points of the plane, those of the original set E in particular, at which all such continuations necessarily coincide. These points constitute a set \underline{E} that we shall call *set of unicity* for the required continuation.

(9.3) *There exists a continuation for which, in addition, every point outside the set of unicity is the center of a circle in which (9.2) is valid with some smaller constant in place of K .*

To see this, we first observe that, for any given point not in \underline{E} , the validity of (9.2) with two distinct values $c \pm \epsilon K$ for $z(x', y')$ when we take x', y' at this point and x'', y'' to lie in the set E , leads to the inequality

$$|z(x'', y'') - c| \leq K \cdot (\rho - \epsilon),$$

from which it follows that the function with the constant value c in the circle of radius ϵ around the given point, and with the value of the original function $z(x, y)$ at each point of E , satisfies (9.2) in the set consisting of E together with this circle. Consequently, applying (9.1) to this function, we see that given any point not in \underline{E} there exists a continuation which is constant in some neighborhood of this point.

This being so, we can cover the complement of \underline{E} by a sequence of such neighborhoods. Denoting by $z_n(x, y)$ the continuation which has a constant value in the n^{th} neighborhood, the function

$$\sum 2^{-n} \cdot z_n(x, y)$$

constitutes a continuation for which in the n^{th} neighborhood (9.2) is valid with the constant K reduced to $(1 - 2^{-n}) \cdot K$. This proves our assertion (9.3).

The reduction of K is not in general possible in the set \underline{E} . In fact, we have the following result.

(9.4) *A point of \underline{E} at a positive distance from the set E is always contained in a segment consisting of points of \underline{E} for every pair of which the relation (9.2) is an equality.*

Moreover, at such a point, the gradient of $z(x, y)$, if it exists, has magnitude K .

We remark, in the first place, that at a point x', y' of the set of unicity the value $z(x', y')$ is the only one which satisfies (9.2) for every x'', y'' of E ; for otherwise we could add this point to the set E with either of two values for $z(x', y')$ and use (9.1) to form two different continuations corresponding to these two values.

Hence, if P is a point of \underline{E} distant $d > 0$ from the set E , then given any positive ϵ there exists a point x'', y'' of E such that, if ρ'' is its distance from P ,

$$|z(x'', y'') - z(P)| > K \cdot (\rho'' - \epsilon);$$

and from this last relation together with (9.2) we deduce that if x', y' is the point on the segment joining P to x'', y'' at distance d from the point P ,

$$|z(x', y') - z(P)| > K \cdot (d - \epsilon).$$

These relations are true for any continuation $z(x, y)$. Making ϵ tend to 0, we see by continuity that there exists a point Q independent of the continuation chosen such that

$$|z(Q) - z(P)| = K \cdot d,$$

the distance of the points P and Q being again equal to d . It is clear moreover that the values of $z(x', y')$ corresponding to the various continuations can only differ by at most K , and hence that the value $z(Q)$ is unique. Finally it is clear from our last equation that (9.2) becomes an equality for every pair of points of the segment PQ , and that this segment lies in the set of unicity \underline{E} .

It now only remains to prove the second part of (9.4).¹⁵ Let R denote the point at distance h from P along a parallel to the x -axis, and let k denote the distance RQ . We write further φ and θ for the angles made by RQ and PQ with the x -axis, and m for the ratio $|z(P) - z(R)|/h$. We have then

$$\begin{aligned} h \cdot m &= |z(P) - z(R)| \geq |z(P) - z(Q)| - |z(R) - z(Q)| \\ &\geq K \cdot d - K \cdot k = K \cdot \{(h^2 + k^2 + 2hk \cos \varphi)^{\frac{1}{2}} - k\}. \end{aligned}$$

Hence, dividing by h and making h tend to 0, we find, since k then tends to d and φ to θ , that the lower limit of m is not less than the value

$$K \cdot \lim_{h \rightarrow 0} \frac{(h^2 + k^2 + 2hk \cos \varphi)^{\frac{1}{2}} - k}{h} = K \cdot \cos \theta.$$

Hence the x -component of the gradient of $z(x, y)$ at P is not less than $K \cdot \cos \theta$, and similarly the y -component is not less than $K \cdot \sin \theta$. This completes the proof.

10. Preliminary reduction

We say that the generalized surface S^* coincides with its track at the point x, y or in the subset B of A if at this point, or almost everywhere in this subset B , the average $M(f)$ defining S^* coincides for every continuous f with the result of substituting for p, q in f the gradient of the track of S^* . We remark that this is certainly the case of any point x, y at which the gradient has the magnitude N of the Lipschitz constant of S^* .

At such a point the gradient has the form $N \cos \alpha, N \sin \alpha$, whence for the function $g(p, q) = N - p \cos \alpha - q \sin \alpha$, we have $M(g) = 0$; and since, in the circle of relevant p, q , this function is non-negative and vanishes only for $p, q = N \cos \alpha, N \sin \alpha$, it follows in the usual way that $M(g) = 0$ for every $g(p, q)$ which vanishes at this point; so that, for any g with the value c at this point, we have $M(g - c) = 0$ i.e. $M(g) = c$, as asserted.

¹⁵ The writer owes the second part of this proof to Mr. J. H. Whiteman.

This being so, we denote by E the boundary of A . The function $z(x, y)$ which defines the track of S^* is a continuation of the type (9.1) for its boundary values in E , the constant K of (9.2) being replaced by N . We form also a second continuation $Z(x, y)$, possibly identical with $z(x, y)$, of the type (9.3).

In the set of unicity \underline{E} we certainly have $z(x, y) = Z(x, y)$. In this set, moreover, since E is closed and of measure zero, almost all points have positive distances from E ; and since the gradient exists almost everywhere, its magnitude is N by (9.4) for almost all points of \underline{E} . Thus S^* coincides with its track in \underline{E} .

We write P, Q for the gradient of $Z(x, y)$ where it exists, and choose $P = Q = 0$ otherwise. We write further

$$p', q' = \epsilon P + (1 - \epsilon)p, \quad \epsilon Q + (1 - \epsilon)q$$

and for any function $g(p, q)$ we define

$$M'(g) = M(g') \quad \text{where} \quad g'(p, q) = g(p', q').$$

With these definitions, the function

$$z'(x, y) = \epsilon Z(x, y) + (1 - \epsilon)z(x, y)$$

has almost everywhere the gradient $M'(p), M'(q)$. Finally we denote by A' a subset of $A - \underline{E}$ consisting of a finite sum of squares in each of which $Z(x, y)$ satisfies a Lipschitz condition with a constant smaller than N . Since every point of $A - \underline{E}$ is the center of such a square by (9.3) we can choose A' so that its measure differs by less than ϵ from that of $A - \underline{E}$.

In A' the magnitude of P, Q has an upper bound N'' less than N , so that for any p, q of magnitude not exceeding N the magnitude of p', q' cannot exceed the value $N' = \epsilon N'' + (1 - \epsilon)N$ which is less than N . Hence for the points x, y of A' , the average $M'(g)$ is wholly independent of the values taken by $g(p, q)$ at the points p, q of magnitude exceeding N' , where N' is less than N .

This being so, we now define the generalized surface S^{**} in the following manner: the track of S^{**} is $z'(x, y)$ in A ; the average is $M'(f)$ at each point x, y of A' ; while in $A - A'$ the surface S^{**} coincides with its track. This generalized surface has the property that its portion in A' has a Lipschitz constant less than N , while the remaining portion is an ordinary surface with the Lipschitz constant N . We shall see that it constitutes an ϵ -approximation to S^* .

To this effect we observe that the relevant pairs of values p', q' and p, q have the difference $\epsilon(P - p), \epsilon(Q - q)$ whose magnitude cannot exceed the number $\eta = 2\epsilon N$; therefore in A' we have

$$|M'(f) - M(f)| = |M(f' - f)| \leq \text{Max} |f(p', q') - f(p, q)| \leq \omega(f; \eta).$$

In \underline{E} on the other hand the two generalized surfaces coincide with a same track, while the remainder of A , the set $A - A' - \underline{E}$, has small measure. From these facts, we easily obtain a relation of the type (8.3), which shows that S^{**} is an ϵ -approximation to S^* , as asserted.

Since this new generalized surface has moreover the same boundary as the

original one, we may substitute it to the latter for the proof of (8.1). In doing so we may at the same time replace the domain A by the finite sum of squares A' , and since the latter can be approximated by a sum of squares without common points it will be sufficient to deal with the case in which A' consists of a single square.

11. Approximation by averages restricted to a finite set

We now make a second reduction which has an interest independent of (8.1). We consider a triangulation of the p, q plane by equilateral triangles of side ϵ , and denote by p_m, q_m , where $m = 1, 2, \dots, n$, the vertices whose magnitude does not exceed N . An average $M(g)$ will be said to be restricted to the p_m, q_m , if it is of the form

$$\sum a_m g(p_m, q_m) \quad \text{where} \quad a_m \geq 0 \quad \text{for each } m \quad \text{and where} \quad \sum a_m = 1.$$

We shall obtain the following result.

(11.1) *A generalized surface S^* with Lipschitz constant less than N has an ϵ -approximation $S^{*'}$ with the same track and with a Lipschitz constant less than N , such that, at each point x, y , the average M' defining $S^{*'}$ is restricted to the p_m, q_m .*

We shall suppose ϵ so small that the triangles with the vertices p_m, q_m enclose all values of p, q which are relevant to the averages $M(f)$ of the definition of S^* . We write in the form

$$(p, q) = \sum b_m \cdot (p_m, q_m)$$

where only the vertices of the triangle containing p, q have non-zero coefficients, the barycentric representation of the point p, q in its triangle or edge of triangle. We define further, for any $g(p, q)$,

$$M'(g) = M(g'), \quad \text{where} \quad g'(p, q) = \sum b_m \cdot g(p_m, q_m);$$

here M is the average occurring in the definition of S^* at any point x, y and this average is defined for the function g' since the latter is continuous, as we easily verify.

Now in the particular case where g is one of the functions p, q or the constant 1, we see at once that $g' = g$, and so $M(g) = M'(g)$. Hence, remembering that the coefficients of the barycentric expression are non-negative, M' is an average with which we can associate the same track as that of S^* to form an $S^{*'}$. Evidently $S^{*'}$ has a Lipschitz constant less than N , so that it only remains to prove that $S^{*'}$ is an ϵ -approximation to S^* . This however follows at once from the relations

$$|M'(g) - M(g)| = |M(g - g')| \leq \sum b_m \cdot \omega(g; \epsilon) = \omega(g; \epsilon).$$

It is convenient to prove also the following more special result.

(11.2) *If the generalized surface S^* has a piece-wise constant gradient, then the approximation $S^{*'}$ satisfying the requirements made in (9.1) can be made to fulfill the additional condition that the finite sum $\sum a_m \cdot g(p_m, q_m)$ which constitutes the*

average $M'(g)$ at the point x, y shall have coefficients a_m which are piece-wise constant in x, y .

To see this, we divide any interval of constancy of the gradient of the track S^* into similar intervals of diameter less than ϵ' , and denote by g_m the function of x, y which is the mean value of a_m in the interval of the division which contains the point x, y . As ϵ' tends to 0, the g_m tend boundedly to the a_m almost everywhere and the convergence is uniform in a subset whose measure is as close as we please to that of the set of definition A . From this it follows that the generalized surface derived from S^* by replacing the a_m by g_m fulfills all our requirements.

12. Reduction to a plane track

The principal part of this section does not concern generalized surfaces. We prove the following theorem.

(12.1) *The track of a Lipschitzian surface with constant less than N is expressible as the limit of a certain track $z_\epsilon(x, y)$ in such a manner that its gradient is almost everywhere the limit of that of $z_\epsilon(x, y)$, where the track defined by $z_\epsilon(x, y)$ satisfies the following additional conditions:*

- (i) *its Lipschitz constant is less than N ,*
- (ii) *outside measure ϵ its gradient is piecewise constant.*

Moreover when these conditions are satisfied, any S^ with a Lipschitz constant less than N and with the original track, has a corresponding ϵ -approximation with Lipschitz constant less than N whose track is $z_\epsilon(x, y)$. Finally the theorem remains true if we restrict throughout the track $z_\epsilon(x, y)$ to have the same boundary as the original track.*

The final remark concerning the boundary is an immediate corollary of the rest of the theorem in view of (9.1), if we apply it in the first place to a domain interior to A whose distance from the boundary of A is positive, and then continue the track obtained suitably to agree with the original track on the boundary of A . We may therefore ignore this part of the assertion, and since we can continue the given track by (9.1) into a square containing A we may suppose that A is such a square.

This being so, let $z(x, y)$ define the original track in A and let $\varphi(x, y)$ be the vector function consisting of the gradient of z , which exists almost everywhere in A . There exists a vector function $\Phi(x, y)$ piecewise constant in A ; we shall suppose it constant in the squares R of side a , such that

$$(12.2) \quad |\varphi - \Phi| < \epsilon^{10} \quad \text{except possibly in measure} < \epsilon^{10}.$$

Let us denote by B the set of those squares R if any for which a subset of R of measure not less than $a^2 \cdot \epsilon^8$ is occupied by points for which the relation $|\varphi - \Phi| < \epsilon^{10}$ is false. Clearly this relation is then false in a subset of B of measure not less than $\epsilon^8 |B|$, so that by (12.2) we must have $|B| < \epsilon^2$.

Consider now any square R not in B . We have then

$$(12.3) \quad |\varphi - \Phi| < \epsilon^{10} \quad \text{except possibly in measure} < a^2 \cdot \epsilon^8 \text{ of } R.$$

We denote by E the linear set contained in the projection of R on the y -axis which consists of the values of y for each of which the relation $|\varphi - \Phi| < \epsilon^{10}$ is false for a corresponding set of x of linear measure not less than $\epsilon^4 a$. It follows from (12.3) that the linear measure of E is less than $\epsilon^4 a$. The same being true when the axes of x and y are interchanged, it follows that we can determine a series of parallels to the two axes, forming a grating over R , in such a way that R is divided into rectangles of sides less than $\epsilon^4 a$ by this grating and that on any segment of this grating in R we have

$$(12.4) \quad |\varphi - \Phi| < \epsilon^{10} \quad \text{except possibly in linear measure} < \epsilon^4 a.$$

Now any point x, y of R can be joined to the center of R by a path C consisting of two portions of segments of the grating and two small segments, the lengths of the latter being less than $\frac{1}{2}\epsilon^4 a$. It follows from (12.4) that on the path C we have

$$(12.5) \quad |\varphi - \Phi| < \epsilon^{10} \quad \text{except possibly in linear measure} < 3\epsilon^4 a.$$

This being so, let $Z(x, y)$ denote the linear function defined in the interior of R to agree with $z(x, y)$ at the center of R and to have the constant gradient Φ . The difference of the functions z and Z at the point x, y cannot exceed in magnitude the integral on C of that of their gradients. Since these gradients are clearly less than N (if ϵ is sufficiently small), it follows from (12.5) that

$$(12.6) \quad |z(x, y) - Z(x, y)| \leq \epsilon^3 \cdot a,$$

provided that ϵ is sufficiently small.

For every R not in B , we now denote by R' the concentric square of side $(1 - 2\epsilon^2) \cdot a$; we denote further by B' the set of the points of the squares R' , and by B'' the set of the points of the boundaries of the corresponding squares R . In the set $B + B' + B''$ we define the function $z_\epsilon(x, y)$ by stipulating that

$$(12.7) \quad z_\epsilon(x, y) = z(x, y) \text{ in } B + B'', \quad z_\epsilon(x, y) = Z(x, y) \text{ in } B'.$$

This function satisfies in $B + B' + B''$ the Lipschitz condition (9.2) with a constant K less than N ; for we need only verify this when the two points x', y' and x'', y'' lie respectively in B' and in $B + B''$. But since the distance ρ of the points then exceeds $\epsilon^2 \cdot a$, we have

$$\begin{aligned} |z_\epsilon(x', y') - z_\epsilon(x'', y'')| &= |Z(x', y') - z(x'', y'')| \\ &\leq \epsilon^3 \cdot a + |z(x', y') - z(x'', y'')| \\ &\leq \epsilon \cdot \rho + K \cdot \rho < N \cdot \rho, \end{aligned}$$

which is the required verification.

This being so, we can continue the function $z_\epsilon(x, y)$ by (9.1) so that it still satisfies the same Lipschitz condition. Hence, observing that any point of A

outside B' is distant at most $\epsilon^2 \cdot a$ from one of coincidence of the functions $z_\epsilon(x, y)$ and $z(x, y)$, we see that the two functions differ by at most $2N \cdot \epsilon^2 a < \epsilon \cdot a$ in $A - B'$ and so by (12.6) throughout A .

Now in B' , except possibly in a subset of measure ϵ^{10} , the gradient of $z_\epsilon(x, y)$ is piecewise constant and differs by less than ϵ^{10} from the gradient of $z(x, y)$. To show that $z_\epsilon(x, y)$ fulfils the requirements of our theorem, it is sufficient to verify that the complement of B' in A has measure less than $\epsilon - \epsilon^{10}$. This is the case since

$$|A - B'| = |B| + \{1 - (1 - 2\epsilon^2)\}^2 \cdot |A - B| \leq \epsilon^2 + 4\epsilon^2 |A|.$$

Now finally let $z_\epsilon(x, y)$ denote any track fulfilling the requirements of the theorem, and let S^* denote a generalized surface with the original track.

We denote by η the lower bound of the numbers such that outside measure η at most

(12.8) the difference of the functions $z(x, y)$, $z_\epsilon(x, y)$ and that of their gradients are both of magnitude at most η .

From our hypotheses, it follows that η tends to 0 with ϵ . We write further A' for the set of x, y in which (12.8) holds, and p', q' for the result of translating p, q by the difference of these gradients. Denoting by M the average defining S^* at any point, we define, for any continuous $g(p, q)$, the average

$$M'(g) = M(g'), \quad \text{where } g'(p, q) = g(p', q')$$

to be at any point of the set A' the average defining $S^{*'}$, and we complete the latter by making it coincide with its track in $A - A'$.

The surface $S^{*'}$ with the track $z_\epsilon(x, y)$ and the average thus defined, is easily verified to be defined consistently and to furnish the required ϵ -approximation.

13. Completion of the proof of (8.1)

The various reductions made possible by the results of the preceding sections show that it is now sufficient to establish (8.1) in the case of a generalized surface S^* defined for the points x, y of a square A , possessing a Lipschitz constant less than N , possessing in view of (12.1)—where we may now choose A to be a square of constancy of the gradient—a plane track; and finally possessing an average at x, y restricted to at most n vectors p_m, q_m , this average being of the form $\sum a_m g(p_m, q_m)$ where the coefficients a_m may be supposed piece-wise constant, and so, by subdivision, absolutely constant, in view of (11.2). With these additional hypotheses, which now cause no real loss of generality, we shall prove our theorem by an induction with respect to the number n .

It will be sufficient to prove that S^* has an ϵ -approximation consisting of a generalized surface $S^{*'}$ formed of a finite number of portions in each of which the corresponding average $M'(g)$ is restricted to at most $n - 1$ of the p_m, q_m .

Moreover we may clearly suppose that A is a triangle instead of a square. Further, we need not insist that the boundary of $S^{*'}$ is to coincide with that of S^* , since this can always be arranged by means of a subsequent modification of

$S^{*'}$ near its boundary, a modification which is made possible by (9.1) and which is, by now, sufficiently clear.

This being so, we may suppose by trivial transformations that the track of S^* is in the x, y plane, and that for a particular value of m , the value $m = 1$ say, we have $q_m = 0$ and $p_m > 0$. We suppose also that the coefficients a_m in the average M are all different from 0, for if this were not the case there would be nothing to prove.

We write k for an index to be summed from $k = 1$ to $k = n - 1$, and we define numbers a_0 and p_0 by the equations

$$a_0 = \sum a_{k+1}, \quad a_0 p_0 = \sum a_{k+1} p_{k+1}.$$

These numbers then satisfy the equations

$$a_0 + a_1 = 1, \quad a_0 p_0 + a_1 p_1 = 0, \quad a_0 > 0, \quad a_1 > 0.$$

We write further $b_k = a_{k+1}/a_0$ and observe that

$$\sum b_k p_{k+1} = p_0, \quad \sum b_k q_{k+1} = 0$$

We now divide the plane into strips parallel to the y -axis whose widths are alternately ϵa_0 and ϵa_1 , and we define a function of x only, $z'(x)$, linear in each strip and taking on the edges separating the strips alternately the values 0 and

$$-\delta = \epsilon a_0 p_0 = -\epsilon a_1 p_1,$$

from which it follows that the gradient of this function in the strips is alternately p_0 and p_1 .

We define the generalized surface $S^{*'}$ to have in the triangle A the track $z'(x, y) = z'(x)$ and the average M' given by

$$M'(g) = \sum b_k g(p_{k+1}, q_{k+1})$$

in the strips where the gradient is p_0 , while at the remaining points the surface coincides with its track i.e. $M'(g) = g(p_1, 0)$. Evidently these conditions are compatible with the definition of our generalized surfaces, since we find that the functions $g = p$ and $g = q$ have the averages p_0 and 0 in the strips where the gradient is p_0 . Since the average M' is restricted in each strip to less than n vectors p_m, q_m it only remains to show that $S^{*'}$ is an ϵ -approximation to S^* .

We observe that on the segment parallel to the x -axis intercepted by a pair of consecutive strips at the height y , the integrals of the two averages $M(f)$ and $M'(f)$ are, for any continuous $f(x, y, z, p, q)$, products of the length of this segment by the expressions of the form

$$\sum a_m f(x, y, 0, p_m, q_m) \quad \text{and} \quad a_1 f(x', y, z', p_1, q_1) + a_0 \sum b_k f(x'', y, z'', p_{k+1}, q_{k+1})$$

where x, x', x'' are intermediate values of x , on account of the mean value theorem of the integral calculus, and where z', z'' are values of z which are small with ϵ . Since $a_m = a_0 b_{m-1}$, these expressions differ from one another by at most $\sum a_m \cdot \omega(f; \eta)$ i.e. by $\omega(f; \eta)$ where η tends to 0 with ϵ .

From this it easily follows that the corresponding double integrals over the triangle A differ by at most $K \cdot \omega(f; \eta)$ and so, that S^{**} is an ϵ -approximation of S^* , as required. This completes the proof.

14. Existence of an attained minimum. Equivalence of generalized and ordinary problem

Let us suppose that we are given the problem of the minimum of one of our functions $F(S)$ in the class of Lipschitzian ordinary surfaces with assigned boundaries. Beside this "ordinary" problem let us consider the "generalized problem" of the minimum of the function $F(S^*)$ in the corresponding class of generalized surfaces S^* .

There exists a sequence of ordinary surfaces $\{S_n\}$ of the class considered, for which the numbers $F(S_n)$ tend to the minimum of the first problem; but, by (8.1), this is also the case in the second problem. The two minima are thus identical, i.e.

(14.1) *The ordinary problem is equivalent to the generalized problem.*

We have established this result in the case in which the surfaces considered are Lipschitzian with unrestricted constants. The argument holds equally when their Lipschitz constants are restricted to be less than or at most equal to some fixed number K . In the latter case we can assert more.

(14.2) *In the class of surfaces with an assigned boundary whose Lipschitz constants do not exceed a fixed number K , the minimum of $F(S)$ is the same for generalized surfaces as for ordinary surfaces and there exists a generalized surface in the class for which this minimum is attained.*

For by (6.1) the minimizing sequence $\{S_n\}$ is compact. Denoting by S^* the limit of a subsequence we evidently have $F(S^*) = \lim F(S_n)$ and this completes the proof.

Let us remark that in the case of a *regular* problem, it is easily seen that the solution S^* must coincide with its track, i.e. reduces to an ordinary surface. This does not however include the whole of what is known of the existence theory for the regular case, since Tonelli¹⁶ obtained a semi-continuity theorem independent of the Lipschitz condition that we have had to assume here. Nevertheless, we shall attempt to show in a subsequent note that even for regular problems the present methods have distinct advantages.

CAPETOWN, SOUTH AFRICA

REFERENCES

- [1] BANACH, S. *Théorie des opérations linéaires*, Monografie Matematyczne (Warszawa, 1932).
- [2] HILBERT, D. *Über das Dirichletsche Prinzip*, Jahresber. der Deutschen Math. Vereinigung, Vol. VIII (1900) pp. 184-188.

¹⁶ Tonelli [9] p. 342.

- [3] LEBESGUE, H. *Sur le problème de Dirichlet*, Rendiconti del Circolo Matematico di Palermo, Vol. 24 (1907) pp. 371–402.
- [4] McSHANE, E. J. *Generalized curves*, Duke Math. Jour., Vol. 6 (1940) pp. 513–536.
- [5] McSHANE, E. J. *Necessary conditions in the generalized-curve problem of the Calculus of Variations*, Ibidem, Vol. 7 (1940) pp. 1–27.
- [6] McSHANE, E. J. *Existence Theorems for Bolza problems in the Calculus of Variations*, Ibidem, Vol. 7 (1940) pp. 28–61.
- [7] McSHANE, E. J. *Extension of range of functions*, Bull. of the Am. Math. Soc., Vol. 40 (1934) pp. 837–842.
- [8] SAKS, S. *Theory of the Integral*, Monografie Matematyczne, Vol. VII (Warszawa, 1937).
- [9] TONELLI, L. *Sur la semi-continuité des intégrales doubles du Calcul des Variations*, Acta Math., Vol. 53 (1929) pp. 325–346.
- [10] WEYL, H. *Die Idee der Riemannschen Fläche*, Mathematische Vorlesungen an der Universität Göttingen, Vol. V (Teubner, Leipzig 1913).
- [11] WEYL, H. *Über die neue Grundlagenkrise der Mathematik*, Math. Zeits., Vol. 10 (1921) pp. 39–79.
- [12] YOUNG, L. C. *On approximation by polygons in the Calculus of Variations*, Proc. of the Royal Soc. (A), Vol. 141 (1933) pp. 325–341.
- [13] YOUNG, L. C. *Generalized Curves and the existence of an attained absolute minimum in the Calculus of Variations*, Comptes Rendus de la Société des Sciences et des Lettres de Varsovie, Classe III, Vol. 30 (1937) pp. 212–234.
- [14] YOUNG, L. C. *Necessary conditions in the Calculus of Variations*, Acta Math., Vol. 69 (1938), pp. 239–258.

FREE LATTICES II

BY PHILIP M. WHITMAN

(Received July 15, 1941)

In the previous paper under this title [14] the author discussed the basic internal structure of free lattices. We assume here a knowledge of the notation used in that paper.¹ In the present paper the previous results are slightly extended and then applied to obtain further properties of free lattices. We study the automorphisms of free lattices (§2), their sublattices (§3), and their order topology (§4), where we find that in the terminology of Birkhoff [3] free lattices are continuous (perhaps vacuously) but not complete.

The author is indebted to Prof. Garrett Birkhoff for inspiration and advice in the preparation of this paper.

1. Introduction

The main definition of [14] will bear repetition: *the free lattice generated by x_1, x_2, \dots, x_n is a lattice generated by them in which there are no laws of equality except those derivable from the postulates for a lattice.* In [14] this lattice was denoted by F_n , since to within isomorphism it depends only on n and not on the symbols x_i . It now seems advisable to replace F_n by $FL(n)$, and also in some cases to specify the generators: $FL(x_1, x_2, \dots, x_n)$, or $FL(X)$ where $X = \{x_1, x_2, \dots, x_n\}$. One may also consider free modular or distributive lattices, denoted $FM(n)$ and $FD(n)$ and defined as are free lattices except that the modular or distributive law ([3] pp. 34, 74) is adjoined to the lattice postulates; cf. Dedekind [7]; Skolem [12]; Birkhoff [1] pp. 450-2, 463-4, [3] pp. 4, 49, 84-85, [4] p. 451; Church [5]. The existence of free algebras in general is shown in Birkhoff [2] p. 441.

The main results of the previous paper [14] may be restated (with slight extensions) as follows, starting with the corollary of §2 and the theorem of §3.

THEOREM 1: *In $FL(n)$, $x_i \leq x_j$ if and only if $i = j$; recursively,*

- (A) $\prod a_i \leq x_j$ if and only if some $a_i \leq x_j$;
- (B) $x_i \leq \sum b_j$ if and only if $x_i \leq$ some b_j ;
- (C) $\sum a_i \leq b$ if and only if every $a_i \leq b$;
- (D) $a \leq \prod b_j$ if and only if $a \leq$ every b_j ;
- (E) $\prod a_i \leq \sum b_j$ if and only if $\prod a_i \leq$ some b_j or some $a_i \leq \sum b_j$.

THEOREM 2: *Of all the polynomials equal in $FL(n)$ to a given polynomial, there is one of shortest length, unique to within commutativity and associativity.*

This is taken as the canonical form. We write $a \equiv b$ for polynomials a and b

¹ Or in [3], plus the notation $L(a)$ = length of a for the number of generators appearing in the polynomial a , counting repetitions.

if one is obtainable from the other by commutativity and associativity. As a criterion for canonicity, we can weaken the conditions of Corollary (18) of [14] slightly:

COROLLARY 1.1: If $\sum a_j$ is

$$(\prod_k a_{1k}) \cup (\prod_k a_{2k}) \cup \dots \cup (\prod_k a_{mk}) \cup x_{i_{m+1}} \cup \dots \cup x_{i_u}$$

then $\sum a_j$ is canonical if and only if (A) for all j and p , $a_{jp} \leq \sum a_k$, and (B) $a_j \leq a_k$ implies $j = k$, and (C) every a_j is canonical. Dually for $\prod a_j$. A generator by itself is canonical.

PROOF: To prove this equivalent to the corollary cited, we need only to show that the failure of (b) in that corollary implies the failure of one of the conditions of Corollary 1.1. Thus suppose that in $FL(n)$, $a_i \leq \sum_{k \neq i} a_k$.

CASE 1: $a_i \equiv x_j$. Then by Theorem 1 B, $x_j \leq a_k$ for some k contrary to (B).

CASE 2: $a_i \equiv \prod_j a_{ij}$. Then by Theorem 1 E, either $\prod_j a_{ij} \leq a_k$ for some k contrary to (B), or

$$a_{ij} \leq \sum_{k \neq i} a_k \leq \sum_{\text{all } k} a_k$$

for some j , contrary to (A).

COROLLARY 1.2: If $\sum a_i = \prod b_j$ in $FL(n)$ and $\sum a_i$ is canonical, then $\prod b_j = b_k$ for some k .

PROOF: This is a companion to Corollary (19) of [14] and is proved similarly.

LEMMA 1.1: In $FL(n)$ for n finite, given the polynomial a and index p then either $x_p \leq a$ or $a \leq \sum_{i \neq p} x_i$, the two being mutually exclusive.

PROOF: This is obvious for $L(a) = 1$; we proceed by induction on $L(a)$.

CASE 1: $a \equiv \prod a_i$. If $a \not\leq \sum_{i \neq p} x_i$, then for all j , $a_j \not\leq \sum_{i \neq p} x_i$, so by induction $x_p \leq a_j$ for all j ; thus $x_p \leq \prod a_j = a$ as desired.

CASE 2: $a \equiv \sum a_i$. If $a \not\leq \sum_{i \neq p} x_i$, then for some k , $a_k \not\leq \sum_{i \neq p} x_i$, so by induction $x_p \leq a_k \leq a$ and thus $x_p \leq a$. Thus at least one of the alternatives holds; they are mutually exclusive since $x_p \leq a \leq \sum_{i \neq p} x_i$ would imply $x_p \leq \sum_{i \neq p} x_i$ and then by Theorem 1, $x_p \leq x_i$ for some i not p which is impossible.

This is a special case of a "splitting" of a lattice [13], a subject which will be discussed in another paper.² An immediate consequence of this lemma is Corollary 3 of §4 of [14], for suppose that for all p , $a \not\leq \sum_{i \neq p} x_i$; then by Lemma 1.1 we have $x_p \leq a$ for all p , so $a \geq \sum_{i=1}^n x_i$; since the latter is the unit element of $FL(n)$, we have $a = \sum_{i=1}^n x_i$, as asserted. The dual of this corollary says that the "points" or "atoms" ([3], p. 10) of $FL(n)$ are the $\prod_{i \neq p} x_i$. We can now strengthen Theorem 3 of [14], not only by extending it farther down in the lattice (which seems relatively uninteresting) but also in the content of its statements about the relation of other elements of the lattice to certain special elements.

THEOREM 3: In $FL(n)$ for n finite, if a is given then either $a = \sum_{i=1}^n x_i$ or

² Soon to be submitted for publication.

$a = \sum_{i \neq p} x_i$ for some p or else for some distinct p and q , $a = b_{pq}$ or $a \leq c_{pq}$, where

$$b_{pq} \equiv \left(\sum_{i \neq p} x_i \right) \cap \left(\sum_{i \neq q} x_i \right)$$

$$c_{pq} \equiv \left(\sum_{i \neq p, q} x_i \right) \cup \sum_{r \neq p, q} \left[\left(\sum_{i \neq p} x_i \right) \cap \left(\sum_{i \neq q} x_i \right) \cap \left(\sum_{i \neq r} x_i \right) \right].$$

Moreover, if $a < \sum_{i \neq p} x_i$, then $a \leq b_{pq}$ for the same p and some q , and likewise if $a < b_{pq}$ then $a \leq c_{pq}$.

PROOF: In view of Corollary 3 of §4 of [14], discussed above, it suffices to prove the last sentence of the theorem; the rest will then follow. Hence we suppose $a < \sum_{i \neq p} x_i$. By Lemma 1.1, if q is given then either $a \leq \sum_{i \neq q} x_i$ or $x_q \leq a$. If the latter held for all q different from p then $\sum_{q \neq p} x_q \leq a \leq \sum_{i \neq p} x_i$ and so $a = \sum_{i \neq p} x_i$ contrary to hypothesis. Hence for some q , $a \leq \sum_{i \neq q} x_i$ as well as $a \leq \sum_{i \neq p} x_i$, and so $a \leq b_{pq}$ for some q other than p , as desired. That $a < b_{pq}$ implies $a \leq c_{pq}$ is proved by similar but more elaborate methods; since we do not use this result we omit the details.

For reference we note certain immediate consequences of Theorem 1; n is not required finite.

COROLLARY 1.3: In $FL(n)$, (A) $\prod a_i \leq \sum y_j$ if and only if for some i , $a_i \leq \sum y_j$, and dually, where the y_j are a non-void subset of the x_i ; (B) if $\sum a_i = x_j$ then $a_k = x_j$ for some k .

PROOF: (A) If $a_i \leq \sum y_j$ for no i , then by Theorem 1 E we would have $\prod a_i \leq y_j$ for some j , and so for some i , $a_i \leq y_j \leq \sum y_k$. (B) By Theorem 1 B, $x_j \leq a_k$ for some k ; $a_i \leq x_j$ for all i , so $a_k = x_j$ by (2) of [14].

2. Automorphisms of $FL(n)$

In some cases an algebra is the free algebra ([3] p. 4; [2] p. 441) for more than one set of generators;³ for instance the free group generated by x and y is the same as that generated by x and $x^{-1}y$. We assert that this is not the case with free lattices. First however we must clarify the meaning of " $\{w_i\}$ is a set of generators." In [14] we said that a set $\{w_i\}$ generates a lattice L if the polynomials in the w_i exhaust L . Now if L consists say of all polynomials in some elements x_i , do we demand that every polynomial a in the x_i should be formally identical with some polynomial in the w_i after substituting the values of the w_i in terms of the x_i ? No, this seems excessive; rather we demand that some polynomial in the w_i should equal a . (As a matter of fact the next theorem holds whichever interpretation is used.) Likewise, as regards automorphisms we consider only their effect on classes of equal elements, not how they may permute polynomials within such a class.

In this section we consider two sets of generators to be the same if their elements are equal by pairs.

DEFINITION 2.1: A set of generators of a lattice L is *redundant* if some member of the set can be omitted without affecting the property that the set generates L .

³ The author is not aware of any precise reference in print, but this is well known. It is mentioned by J. Dyer-Bennet, thesis, *A free algebra with three operations*, unpublished, Harvard University, 1940.

THEOREM 4: *$FL(n)$ is the free lattice on a unique set X of generators; any set of generators must contain X ; and X is the only irredundant set of generators.*

PROOF: First we observe that $FL(n)$ could not be the free lattice on a redundant set of generators, for if it were then by Definition 2.1 one of the generators would equal a polynomial in the others, which is not true in a free lattice. Now let $FL(n)$ be $FL(x_1, \dots, x_n)$; that is, it consists of all lattice polynomials in the x_i , subject to the laws stated in Theorem 1; and suppose that it is also a lattice generated by w_1, \dots, w_m . We assert that given k , then $w_i = x_k$ for some i . For suppose not: for some k and all i , $w_i \neq x_k$; then we may proceed by induction on $L(a)$ where a is any polynomial in the w_i . Suppose $a \neq x_k$ for all a with $L(a) < h$. Then suppose $a = x_k$ and $L(a) = h$; $a \equiv \sum a_i$ or dually. By Corollary 1.3 B, $a_i = x_k$ for some i , since a is indirectly a polynomial in the x_i , contrary to hypothesis of induction. Thus by induction we would have that no polynomial in the w_i equals x_k so the w_i could not be a set of generators. Hence $w_i = x_k$ for some i , given k , proving the second clause of the theorem. If moreover $\{w_i\}$ is irredundant then $\{w_i\}$ must coincide with $\{x_i\}$ to within equality, which proves the last clause. Then the first part of the proof proves the first clause. Q.E.D.

This states only that a different irredundant set of generators cannot generate all of $FL(n)$; as we shall see in the next section, they may generate a sublattice isomorphic to it. Another consequence is this: any automorphism of a lattice is by definition determined by the images of the generators, so by Theorem 3 an automorphism of $FL(n)$ must merely permute the x_i among themselves. On the other hand, by the symmetry of the postulates any such permutation determines an automorphism. Hence

COROLLARY 2.1: *The group of automorphisms of $FL(n)$ is the symmetric group on n symbols.*

3. Sublattices of $FL(n)$

We observe first that in free lattices the relations between polynomials in x_1, \dots, x_k are independent of the presence or absence of other generators:

LEMMA 3.1: *If $f(x_1, \dots, x_k) \leq g(x_1, \dots, x_k)$ for polynomials f and g in $FL(n)$, then this relation is also true in $FL(k)$, and conversely.*

PROOF: By Theorem 1 the conditions in the two cases are the same.

LEMMA 3.2: *If $\{x_1, \dots, x_r\}$, $\{y_1, \dots, y_s\}$, and $\{z_1, \dots, z_t\}$ are disjoint subsets of the generators of $FL(n)$, and*

$$f(x_1, \dots, x_r, y_1, \dots, y_s) \leq g(x_1, \dots, x_r, z_1, \dots, z_t)$$

for polynomials f and g in $FL(n)$, then also

$$f(x_1, \dots, x_r, 1, \dots, 1) \leq g(x_1, \dots, x_r, 0, \dots, 0)$$

in $FL(n)$.⁴

⁴ 0 and 1 stand for the zero and unit elements of the lattice. If not present, they may for the purposes of this lemma be temporarily adjoined; cf. MacNeille [9] p. 443. They will nearly always be absorbed in reducing the new polynomials to simpler form.

PROOF: This is trivial for $L(f) + L(g) = 2$. Denoting the results of the substitutions by bars, we make an induction on $L(f) + L(g)$. If $g = \prod g_i$, then by Theorem 1 $f \leq g_i$ for all i , so by induction $\bar{f} \leq \bar{g}_i$ for all i , and hence $\bar{f} \leq \prod \bar{g}_i = \bar{g}$. If $g = \sum g_i$ and $f = \prod f_i$, then $f \leq g_i$ for some i , or $f_i \leq g$ for some i , so $\bar{f} \leq \bar{g}_i$ or $\bar{f}_i \leq \bar{g}$, and thus $\bar{f} \leq \bar{g}$. The other cases are similar. Q.E.D.

Let us now consider the sublattices of $FL(x_1, \dots, x_n)$ for n fixed. In particular, we inquire which of these sublattices are again free lattices on suitable generators. Since a sublattice is often most easily designated by listing its generators, we shall find it convenient to make the following

DEFINITION 3.1: *The set $\{u_i\}$ is free if the sublattice generated by it is the free lattice with these generators.*

The question then becomes: what conditions on the u_i will insure that they form a free set? In Theorem 5 it appears that the answer is simply that with respect to their joins and meets they shall behave as do the x_i . The situation in free lattices is thus unlike that in free groups, where every subgroup is again free.⁵

THEOREM 5: *A subset $U = \{u_i\}$ of the elements of $FL(n)$ is free if and only if $u_j \leq \sum_{i \in S} u_i$ and its dual (where S is a finite set of indices) each imply $j \in S$.*

PROOF: We wish to show these conditions necessary and sufficient that in the given lattice $FL(n) = FL(X)$, the sublattice generated by U is $FL(U)$, or equivalently that this sublattice is isomorphic to $FL(Y) = FL(y_1, \dots, y_k)$ where k is the cardinal power of the set U . Thus we wish to show that the conditions of the theorem are necessary and sufficient in order that for all polynomials f and g ,

$$f(u_{i_1}, \dots, u_{i_j}) \leq g(u_{i'_1}, \dots, u_{i'_j})$$

in $FL(X)$ if and only if

$$f(y_{i_1}, \dots, y_{i_j}) \leq g(y_{i'_1}, \dots, y_{i'_j})$$

in $FL(Y)$, or (more briefly) in order that $f(U) \leq g(U)$ in $FL(X)$ if and only if $f(Y) \leq g(Y)$ in $FL(Y)$. The necessity of the given conditions is apparent from the fact that by Theorem 1 the y_i behave in this manner. For sufficiency we assume the conditions of the theorem and prove the isomorphism by induction on $L(f(Y)) + L(g(Y))$. If $L(f(Y)) + L(g(Y)) = 2$ then $L(f(Y)) = L(g(Y)) = 1$. By hypothesis $u_i \leq u_j$ if and only if $i = j$, which is true if and only if $y_i \leq y_j$, as desired. Proceeding by induction,

CASE 1: $f(Y) = y_j$. Now $y_j \leq g(Y)$ implies $u_j \leq g(U)$ because the lattice generated by U is a homomorphic image of $FL(Y)$ by [2], Theorem 9 on p. 441. Conversely if the latter holds then the former does, for if not then by Lemmas 1.1 and 3.1 applied to the finite set R consisting of the y_i which appear in either $f(Y)$ or $g(Y)$,

⁵ Nielsen [10], Schreier [11]. Levi [8] gives a new proof. The falsity for free lattices is shown by taking any two-element sublattice.

$$g(Y) \leq \sum_{i \in R-j} y_i$$

in $FL(y_i; i \in R)$ and so also in $FL(Y)$. But then

$$u_j \leq g(U) \leq \sum_{i \in R-j} u_i$$

contrary to hypothesis.

CASE 2: $g(Y) = y_j$. Dual of case 1.

CASE 3: $f(Y) = \sum f_i(Y)$. Then $\sum f_i(U) \leq g(U)$ implies by Theorem 1 that $f_i(U) \leq g(U)$ for all i , so by induction $f_i(Y) \leq g(Y)$ for all i , whence $f(Y) \leq g(Y)$, while the converse follows as at the beginning of Case 1.

CASE 4: $g(Y) = \prod g_i(Y)$. Dual of case 3.

CASE 5: $f(Y) = \prod f_i(Y)$ and $g(Y) = \sum g_i(Y)$. Then $f(U) \leq g(U)$ implies by Theorem 1 that $f_i(U) \leq g(U)$ for some i or else $f(U) \leq g_i(U)$ for some i , and by induction we again get $f(Y) \leq g(Y)$, and conversely as above. Q.E.D.

In particular we can verify that in $FL(3)$ these conditions are satisfied by $x_1 \cup (x_2 \cap x_3)$, $x_2 \cup (x_1 \cap x_3)$, and $x_3 \cup (x_1 \cap x_2)$, or still better,

LEMMA 3.3: *In $FL(3)$ the following are a free set:*

$$u_1 = [x_1 \cap (x_2 \cup x_3)] \cup [x_2 \cap (x_1 \cup x_3)]$$

$$u_2 = [x_1 \cap (x_2 \cup x_3)] \cup [x_3 \cap (x_1 \cup x_2)]$$

$$u_3 = [x_1 \cup (x_2 \cap x_3)] \cap [x_2 \cup (x_1 \cap x_3)]$$

$$u_4 = [x_1 \cup (x_2 \cap x_3)] \cap [x_3 \cup (x_1 \cap x_2)].$$

The proof is a matter of straightforward verification of the conditions of Theorem 5, using Theorem 1. We give one example: $u_4 \not\leq u_1 \cup u_2 \cup u_3$. For otherwise one of the following must hold: (i) $u_4 \leq u_1$; (ii) $u_4 \leq u_2$; (iii) $u_4 \leq u_3$; (iv) $x_1 \cup (x_2 \cap x_3) \leq u_1 \cup u_2 \cup u_3$; (v) $x_3 \cup (x_1 \cap x_2) \leq u_1 \cup u_2 \cup u_3$, by Theorem 1 E. But if (i) held, then either $u_4 \leq x_1 \cap (x_2 \cup x_3)$ or $u_4 \leq x_2 \cap (x_1 \cup x_3)$ or $x_1 \cup (x_2 \cap x_3) \leq u_1$ or $x_3 \cup (x_1 \cap x_2) \leq u_1$, which are contradicted respectively by the facts that $u_4 \not\leq x_1$, $u_4 \not\leq x_2$, $x_1 \not\leq u_1$, and $x_3 \not\leq u_1$, so (i) is impossible. Similarly so is (ii). If (iii) held, then $u_4 \leq x_2 \cup (x_1 \cap x_3)$; thus $u_4 \leq x_2$ or $u_4 \leq x_1 \cap x_3$ or $x_1 \cup (x_2 \cap x_3) \leq x_2 \cup (x_1 \cap x_3)$ or $x_3 \cup (x_1 \cap x_2) \leq x_2 \cup (x_1 \cap x_3)$. These are respectively contradicted by the facts that $u_4 \not\leq x_2$, $u_4 \not\leq x_1 \cap x_3$, $x_1 \not\leq x_2 \cup (x_1 \cap x_3)$, and $x_3 \not\leq x_2 \cup (x_1 \cap x_3)$, so (iii) is false. (iv) and (v) are false since $x_1 \not\leq u_1 \cup u_2 \cup u_3$ and $x_3 \not\leq u_1 \cup u_2 \cup u_3$. Thus the typical condition chosen holds; the others are similar. Q.E.D.

THEOREM 6: *$FL(3)$ has $FL(n)$ as a sublattice, for any finite or countable n .*

PROOF: By Lemma 3.3, $FL(3)$ has a free set $\{u_1, u_2, u_3, u_4\}$. We proceed by induction on the number of elements in a free set $\{u_i\}$. Suppose $\{u_1, u_2, \dots, u_k\}$ is a free set of elements of $FL(3)$. By Lemma 3.1, $\{u_{k-2}, u_{k-1}, u_k\}$ is free, so by Lemma 3.3 there is a free set $\{v_1, v_2, v_3, v_4\}$ of polynomials in u_{k-2}, u_{k-1} , and u_k . Then

$$\{u_1, u_2, \dots, u_{k-3}, v_1, v_2, v_3, v_4\}$$

is free, for we may readily verify the conditions of Theorem 5:

$$\text{CONDITION 1: } u_1 \not\leq u_2 \cup u_3 \cup \dots \cup u_{k-3} \cup v_1 \cup v_2 \cup v_3 \cup v_4.$$

For otherwise

$$u_1 \leq u_2 \cup \dots \cup u_{k-3} \cup v_1 \cup v_2 \cup v_3 \cup v_4 \leq u_2 \cup \dots \cup u_k$$

contrary to hypothesis that $\{u_1, \dots, u_k\}$ was free.

$$\text{CONDITION 2: } v_4 \not\leq u_1 \cup u_2 \cup \dots \cup u_{k-3} \cup v_1 \cup v_2 \cup v_3.$$

For otherwise (since v_4 is a polynomial in u_{k-2} , u_{k-1} , and u_k alone) by Lemma 3.2, $v_4 \leq v_1 \cup v_2 \cup v_3$ contrary to the choice of the v_i .

The other conditions follow by symmetry and duality. Thus the conditions for induction are fulfilled and we obtain a countable set of u_i . This whole set together is free, for the conditions of Theorem 5 involve only finite sets and so hold by construction. Q.E.D.

COROLLARY 3.1: *For $n \geq 3$ and any finite or countable k , $FL(n)$ has $FL(k)$ as a sublattice.*

THEOREM 7: *$FL(n)$ has no sublattice isomorphic to the free modular or distributive lattice on more than two generators.*

PROOF: Otherwise $FL(n)$ would have $FM(3)$ or $FD(3)$ as a sublattice, for if say it had say $FM(4)$ as a sublattice, then the sublattice generated by the first three of these four generators would satisfy the definition of free modular lattice (this argument assumes well-ordering in the infinite case); a similar argument would have sufficed for Lemma 3.1. But it is known ([7] p. 246; [3] pp. 49, 84) that the generators of $FM(3)$ and $FD(3)$ satisfy the conditions of Theorem 5, so the sublattice they generate would be $FL(3)$ rather than $FM(3)$ or $FD(3)$.

4. Intrinsic topology of $FL(n)$

In a lattice each pair of elements, a and b , must have a least upper and greatest lower bound ([3] p. 16, [14] p. 325). By the associative law this must also hold for any finite number of elements, but it may or may not hold for an infinite number (e.g., the set of all integers is a lattice under the customary linear ordering, but there is no greatest; on the other hand any class of subclasses of a fixed class does have a least upper bound under set inclusion). In case the least upper bound or greatest lower bound of a set $\{a_i\}$ exists we may denote it $\sup \{a_i\}$ or $\inf \{a_i\}$ respectively ([3] p. 16).

DEFINITION 4.1 ([3] p. 27): In a lattice, $\{a_i\}$ is said to (*o*)-converge to a if sequences $\{u_i\}$ and $\{v_i\}$ exist such that

$$u_i \leq u_{i+1} \leq a_{i+1} \leq v_{i+1} \leq v_i$$

for all i , and $\sup \{u_i\} = \inf \{v_i\} = a$. Then we write $a_i \rightarrow a$, $u_i \uparrow a$, and $v_i \downarrow a$.

In terms of this order convergence, one may study the topology of a lattice—

the intrinsic topology, as Birkhoff calls it, since it is defined in terms of the inclusion relation without external considerations. For instance, the lattice operations are said to be *continuous*⁶ if $a_i \rightarrow a$ and $b_i \rightarrow b$ imply $a_i \cup b_i \rightarrow a \cup b$ and dually. For this, it is sufficient ([3] p. 30) that $a_i \downarrow a$ imply $a_i \cup b \downarrow a \cup b$ and dually. If $\sup \{a_i\}$ and $\inf \{a_i\}$ exist for every set $\{a_i\}$ then the lattice is said to be *complete*.

We have already (§4 of [14]) obtained a simple topological result for $FL(n)$, in finding that in certain cases one element covers another; thus $FL(n)$ is in a sense not everywhere dense. Moreover, in Theorem 9 below we find that $FL(n)$ is not complete—we exhibit sets without suprema. This leaves open the question of the existence of an infinite ascending chain with a supremum; if there is none then Theorem 8 is vacuous.

THEOREM 8: $FL(n)$ is continuous.

PROOF: Suppose $a_i \downarrow a$. By [3] (p. 30) it suffices to show $a_i \cup b \downarrow a \cup b$. But $a \cup b$ is obviously a lower bound to the $a_i \cup b$; it remains to be shown greatest—that is, that if c is any lower bound to the $a_i \cup b$ then $c \leq a \cup b$. But if this should fail—if $c \not\leq a \cup b$ —then $c \cup a \cup b > a \cup b$ and $c \cup a \cup b$ is again a lower bound, so it suffices to show by induction on $L(c)$ that it is impossible to have $c > a \cup b$ and c a lower bound to the $a_i \cup b$.

CASE 1: $c \equiv \sum_{j=1}^m c_j$, where we may suppose that no c_j is a join. Then $c \leq a_i \cup b$ for all i , so $c_j \leq a_i \cup b$ for all i and j . Hence by Theorem 1, given i and j , either (1) $c_j \leq a_i$, or (2) $c_j \leq b$, or (3) $c_j \equiv \prod_k c_{jk}$ and for some k , $c_{jk} \leq a_i \cup b$, the possibility (3) being excluded if c_j is a generator. If (3) holds for some j and infinitely many i , it holds for that j and infinitely many i and some fixed k , since m is finite. By monotonicity of the a_i , $c_{jk} \cup \sum_{i \neq j} c_i$ is a shorter lower bound than c and contains $a \cup b$, contrary to hypothesis of induction. Thus, given j , it must be that (1) holds for infinitely many i or (2) does; perhaps both. In the former case $c_j \leq a$, in the latter $c_j \leq b$, so $c \equiv \sum c_j \leq a \cup b$ contrary to hypothesis that $c > a \cup b$.

CASE 2: $c \equiv \prod_{j=1}^m c_j$. Then $\prod c_j \leq a_i \cup b$ for all i , so given i either (1) $c \leq a_i$, or (2) $c \leq b$, or (3) for some j , $c_j \leq a_i \cup b$. If (3) holds for infinitely many i , then (since m is finite) it holds for some fixed j and infinitely many i , and hence for that j and all i , and c_j is a shorter lower bound, $c_j > a \cup b$, contrary to induction. Otherwise (1) or (2) holds for infinitely many i , so $c \leq a$ or $c \leq b$, contrary to $c > a \cup b$.

CASE 3: c is a generator. This is included in Case 2, with (3) impossible. This starts the induction. Thus the theorem holds.

This leaves open the question of whether there are any infinite sequences which have limits. We proceed to show that at any rate in $FL(3)$ not every infinite sequence has a limit. For our horrible example we consider the elements

⁶ Not to be confused with the use by some authors of “continuous” for what we call “complete,” nor with the use of “continuous” in connection with a metric in a lattice ([3], p. 43).

which Birkhoff used⁷ to show that $FL(3)$ has an infinite number of distinct elements.

Let $t_1 \equiv x_1$, $t_{6i+2k} \equiv t_{6i+2k-1} \cap x_{2k}$, $t_{6i+2k+1} \equiv t_{6i+2k} \cup x_{2k+1}$ for $i = 0, 1, 2, \dots$, where the subscripts on x are taken modulo 3. Thus for instance

$$t_{6i+7} \equiv (((((t_{6i+1} \cap x_2) \cup x_3) \cap x_1) \cup x_2) \cap x_3) \cup x_1).$$

By Theorem 1, $t_1 \leq t_7$, and by induction $t_{6h+k} \leq t_{6i+k}$ in $FL(3)$, if $h < i$. Birkhoff's example just cited shows that equality cannot hold, as could also be shown by Theorem 1 and induction, so $t_{6h+k} < t_{6i+k}$ in $FL(3)$ if $h < i$. We have thus six infinite strictly monotone increasing chains: t_1, t_7, t_{13}, \dots ; t_2, t_8, t_{14}, \dots , etc.

LEMMA 4.1: *If one of the infinite chains t_{6i+k} (k fixed) has a least upper bound, then so does each, and if $\sup_i \{t_{6i+k}\} = s_k$, then $s_{2k} = s_{2k-1} \cap x_{2k}$ and $s_{2k+1} = s_{2k} \cup x_{2k+1}$.*

NOTE: The subscripts on s may be taken modulo 6, on x modulo 3. We do not as yet assert that $s_{2k-1} \cap x_{2k}$ gives the canonical form of s_{2k} .

PROOF: By Theorem 8, if s_{2k-1} exists then

$$\begin{aligned} s_{2k} &\equiv \sup_i \{t_{6i+2k}\} \equiv \sup_i \{t_{6i+2k-1} \cap x_{2k}\} \\ &= \sup_i \{t_{6i+2k-1}\} \cap \sup_i x_{2k} \\ &\equiv s_{2k-1} \cap x_{2k} \end{aligned}$$

so that s_{2k} exists and equals $s_{2k-1} \cap x_{2k}$. Likewise $s_{2k+1} = s_{2k} \cup x_{2k+1}$ and so if any s exists, then by repetition all do. Q.E.D.

LEMMA 4.2: *If the s_k are in canonical form, then the equalities in Lemma 4.1 become identities.*

PROOF: PART (I): $s_{2k+1} = s_{2k} \cup x_{2k+1}$.

CASE 1: $s_{2k+1} \equiv \prod r_i$. Then either $s_{2k+1} = s_{2k}$ or $s_{2k+1} = x_{2k+1}$, by Corollary 1.2. But the second is contradicted by $x_{2k+1} \leq t_{6i+2k+1} < s_{2k+1}$, and if the former held, then by Lemma 4.1,

$$x_{2k+1} < s_{2k+1} = s_{2k} \leq x_{2k}$$

whereas $x_{2k+1} \not\leq x_{2k}$ in $FL(3)$. Thus Case 1 is impossible.

CASE 2: $s_{2k+1} \equiv \sum r_j$. By Corollary (19) of [14] and by (I), we have

$$(1) \quad \text{given } j, \text{ either } r_j \leq s_{2k} \text{ or } r_j \leq x_{2k+1}.$$

Now $x_{2k+1} \leq s_{2k+1}$, so by Theorem 1 B, $x_{2k+1} \leq r_j$ for some j ; say

$$(2) \quad x_{2k+1} \leq r_1.$$

But as in Case 1,

$$(3) \quad x_{2k+1} \not\leq s_{2k}.$$

⁷ [2] pp. 451-452. The following typographical changes are needed on p. 452: line 5, subscript should be $i + 1$; lines 7-8, interchange meet and join symbols.

By (2) and (3), $r_1 \not\leq s_{2k}$, so by (1) and (2), $r_1 = x_{2k+1}$. By canonicity and (1), we have

$$(4) \quad r_1 = x_{2k+1} \quad \text{and} \quad r_j \leq s_{2k} \quad \text{for} \quad j > 1.$$

By Lemma 4.1,

$$s_{2k-1} \cap x_{2k} = s_{2k} \leq s_{2k+1} \equiv \sum r_j,$$

so by Theorem 1, either

$$(5) \quad s_{2k} \leq r_i \quad \text{for some } i, \quad \text{or}$$

$$(6) \quad s_{2k-1} \leq s_{2k+1} \quad \text{or}$$

$$(7) \quad x_{2k} \leq s_{2k+1}.$$

If (7) holds, then $x_{2k} \cup x_{2k+1} \leq s_{2k+1}$, which is impossible in $FL(3)$ by the covering results stated in Theorem 3 since s_{2k+1} is the limit of a strictly monotone increasing sequence. If (6) holds then a similar contradiction is reached: $x_{2k-1} \cup x_{2k+1} \leq s_{2k+1}$. Hence (5) holds. But $s_{2k} \not\leq x_{2k+1}$, for

$$x_{2k-1} \cap x_{2k} \leq (s_{2k-2} \cup x_{2k-1}) \cap x_{2k} = s_{2k-1} \cap x_{2k} = s_{2k}$$

whereas $x_{2k-1} \cap x_{2k} \not\leq x_{2k+1}$. Hence in (5) $i > 1$ and by (4), $r_i = s_{2k}$. By canonicity,

$$s_{2k+1} \equiv r_1 \cup r_2 \equiv x_{2k+1} \cup s_{2k}.$$

CASE 3: $s_{2k} \equiv x_i$. Like Case 1.

PART (II): $s_{2k} = s_{2k-1} \cap x_{2k}$.

CASE 1: $s_{2k} \equiv \sum r_j$. By Corollary 1.2, either

$$(1) \quad s_{2k} = s_{2k-1} \quad \text{or}$$

$$(2) \quad s_{2k} = x_{2k}.$$

If (1) holds, then

$$x_{2k-1} \leq s_{2k-1} = s_{2k} \leq x_{2k}$$

whereas $x_{2k-1} \not\leq x_{2k}$. If (2) holds, then

$$x_{2k} \cup x_{2k+1} \leq s_{2k} \cup x_{2k+1} = s_{2k+1}$$

contrary to Theorem 3.

CASE 2: $s_{2k} \equiv \prod r_j$. By Corollary (19) of [14], given j , either

$$(1) \quad s_{2k-1} \leq r_j \quad \text{or}$$

$$(2) \quad x_{2k} \leq r_j.$$

Also $s_{2k} \leq x_{2k}$, so for some j , $r_j \leq x_{2k}$; say $j = 1$:

$$(3) \quad r_1 \leq x_{2k}.$$

Then $s_{2k-1} \not\leq r_1$ (otherwise $x_{2k-1} \leq s_{2k-1} \leq r_1 \leq x_{2k}$, whereas $x_{2k-1} \not\leq x_{2k}$) so by (2), $x_{2k} \leq r_1$, and by (3) and canonicity we have

$$(4) \quad x_{2k} \equiv r_1 \quad \text{and} \quad s_{2k-1} \leq r_j \quad \text{for} \quad j > 1.$$

Also

$$\prod r_i \equiv s_{2k} = s_{2k-1} \cap x_{2k} \leq s_{2k-1} = s_{2k-2} \cup x_{2k-1},$$

so either

$$(5) \quad s_{2k} \leq s_{2k-2} \quad \text{or}$$

$$(6) \quad s_{2k} \leq x_{2k-1} \quad \text{or}$$

$$(7) \quad r_i \leq s_{2k-1} \quad \text{for some } i.$$

If (5) holds, then $s_{2k} \leq x_{2k-2} \cap x_{2k}$, and if (6) holds then $s_{2k} \leq x_{2k-1} \cap x_{2k}$, each contrary to Theorem 3. Hence (7) holds, and by (4) and canonicity,

$$s_{2k} \equiv x_{2k} \cap s_{2k-1}.$$

CASE 3: $s_{2k} \equiv x_i$. Like Case 1.

Thus Lemma 4.2 is proven.

LEMMA 4.3: *FL(3) is not complete.*

PROOF: The assumption that every set has a least upper bound leads via Lemmas 4.1 and 4.2 (applied six times) to the conclusion that

$$s_1 = (((((s_1 \cap x_2) \cup x_3) \cap x_1) \cup x_2) \cap x_3) \cup x_1$$

and that both sides of this equation are in canonical form, which is impossible since they are not of the same length.

THEOREM 9: *FL(n) is not complete*⁸, for $n \geq 3$.

PROOF: *FL(n)* has the infinite chain $\{t_{6i+1}\}$, of which each element is a polynomial in x_1, x_2, x_3 alone, not involving x_4, \dots, x_n . By⁹ Lemma 3.2, the least upper bound of this chain (if it exists) is a polynomial in x_1, x_2, x_3 alone, and so by Lemma 3.1 must be the least upper bound in *FL(3)*, contrary to the proof of Lemma 4.3.

5. *FM(4)*

Birkhoff has shown ([1] p. 463-464) that the free modular lattice *FM(4)* has infinitely many distinct elements. A slightly stronger result, again due to Birkhoff¹⁰ is this:

THEOREM 10: *FM(4) has an infinite chain of distinct elements.*

PROOF: Let A be the free Abelian group generated by $x_1, x_2, \dots, x_n, \dots$. For $k = 0, 1, 2, \dots$ let $S_1 = [x_{2k}]$, the subgroup of A generated by the x_{2k} , $S_2 = [x_{2k} + x_{2k+1}]$, $S_3 = [x_{2k+1}]$, $S_4 = [x_{2k+1} + x_{2k+2}]$. If we set $B = [S_1, S_2,$

⁸ In fact, not even conditionally sigma-complete ([3] p. 29); we have exhibited a bounded countable set without least upper bound. Note: While n was required finite in Theorem 3, the method of applying that theorem here avoids such a restriction in the present theorem.

⁹ The adjunction of 1 and 0 possibly required for that lemma causes no trouble, for (using Lemma 3.1) we may apply Lemma 3.2 to the lattice generated by the first three x 's and those involved in an alleged least upper bound; since this set is finite, 1 already exists for this sublattice.

¹⁰ Who communicated this section to the author by letter, March 7, 1940.

S_3, S_4], then the set of all subgroups of B is a modular lattice ([3] p. 35). Let $T_1 = S_1, T_{4n+1} = T_{4n} \cap S_1, T_{4n+2} = T_{4n+1} \cup S_2, T_{4n+3} = T_{4n+2} \cap S_3, T_{4n+4} = T_{4n+3} \cup S_4$. Then

$$\begin{aligned} T_{4n} &= [x_{2n+2k-1}, x_{2n+2k}, x_{2k+1} + x_{2k+2}] \\ T_{4n+1} &= [x_{2n+2k}] \\ T_{4n+2} &= [x_{2n+2k}, x_{2n+2k+1}, x_{2k} + x_{2k+1}] \\ T_{4n+3} &= [x_{2n+2k+1}], \end{aligned}$$

whence $T_4 > T_8 > T_{12} > \dots$. Thus one modular lattice with four generators has an infinite chain, hence so does $FM(4)$. Q.E.D.

6. Unsolved problems

As far as the author is aware, the following problems suggested by or related to the present paper remain open. (1) Solve the word problem for free modular lattices ([3] p. 146). This seems to be difficult. (2) Extend Theorem 3, and [14] §4. Perhaps not difficult. (3) Does *some* infinite set in $FL(n)$ order-converge (§4)? We may call a set of elements $\{a_1, a_2, \dots\}$ a *fence* below a if $c < a$ implies $c \leq a_i$ for some i . As we saw in Theorem 3 the $\sum_{i \neq p} x_i$ form a fence below the unit in $FL(n)$; etc. Is it perhaps true that in $FL(n)$ *every* element has a finite upper and lower fence? This would settle the first part of (3). On the other hand if some suprema exist, are there real cases like the vacuous one of Lemma 4.2 where canonical forms carry over? Do there exist infinite connected chains in $FL(n)$ —ones in which no further elements can be interpolated? (4) Study the free complete lattice. Study the completion of $FL(n)$ by cuts (MacNeille [9] p. 443; [3] p. 27). (5) Determine the finite sublattices of $FL(n)$; cf. Theorem 7.

HARVARD UNIVERSITY AND THE UNIVERSITY OF PENNSYLVANIA

REFERENCES

- G. BIRKHOFF. [1]: *On the combination of subalgebras*, Proc. Camb. Phil. Soc. 29 (1933) pp. 441–464. [2]: *On the structure of abstract algebras*, *ibid.*, 31 (1935) pp. 433–454. [3]: *Lattice Theory*, Amer. Math. Soc. Colloq. Pub. XXV, New York 1940. [4]: *Rings of Sets*, Duke J. 3 (1937) pp. 443–454.
- R. CHURCH. [5]: *Numerical analysis of certain free distributive structures*, Duke J. 6 (1940) pp. 732–734.
- R. DEDEKIND. [6]: *Über Zerlegungen von Zahlen durch ihre grosssten gemeinsam Teiler*, Festschrift Techn. Hoch. Braunschweig (1897), Gesammelte Werke II pp. 103–148. [7]: *Über die von drei Moduln erzeugte Dualgruppe*, Math. Annalen 53 (1900) pp. 371–403, Ges. Werke II pp. 236–271.
- F. LEVI. [8]: *Über die Untergruppen freier Gruppen*, Math. Zeit. 32 (1930) p. 315ff.
- H. M. MACNEILLE. [9]: *Partially ordered sets*, Trans. Am. Math. Soc. 42 (1937) pp. 416–460.
- J. NIELSEN. [10]: *Om regning med ikke-kommutative faktorer*, Matematisk Tidsskrift B, 1921, pp. 77–94.
- O. SCHREIER. [11]: *Die Untergruppen der freien Gruppen*, Abh. aus den Math. Seminar Hamburg 5 (1927) pp. 161–183.
- T. SKOLEM. [12]: *Über gewisse Verbände oder Lattices*, Avhandlingar utgitt av det Norske Videnskaps Akademi (Mat.-Naturv. Klasse) 1936 no. 7.
- P. M. WHITMAN. [13]: *Abstract*, Bull. Am. Math. Soc. 46 (1940) p. 760. [14]: *Free Lattices*, Annals of Math. (2) 42 (1941) pp. 325–330.

TOPOLOGICAL METHODS FOR THE CONSTRUCTION OF TENSOR FUNCTIONS

By NORMAN E. STEENROD

(Received July 7, 1941)

1. Introduction

It is a well-known theorem that an orientable manifold M admits a continuous field of non-zero tangent vectors if and only if the Euler number of M is zero. Stiefel [5]¹ has generalized this result to the case of any finite number of fields independent at each point, obtaining necessary conditions in the form of the vanishing of certain cohomology classes of M .

From these results, there is every reason to expect a general theory connecting the existence of tensor functions of various types over a manifold M with the topological structure of M . It is our purpose in this paper to give the foundations and first theorems of such a theory.

Briefly, it is shown that the set M' of point tensors (of a fixed order and weight) over the differentiable manifold M is a differentiable manifold forming, in a natural way, a fibre bundle over M in the sense of Whitney [6]. A tensor function attaches to each point of M a point of the fibre in M' over it. The nature of a tensor function is restricted by specifying that its values lie in a submanifold M'' of M' . If M'' is likewise a fibre bundle, then there is a characteristic cohomology class in M attached to M'' . The vanishing of this class proves to be a necessary (and sometimes sufficient) condition for the existence of the prescribed tensor function. As an application, a direct and simple proof is given that any separable manifold admits a Riemann metric.²

It is worth noting here that the characteristic cohomology class belongs to a new type of cohomology group: one based on local coefficient groups connected by local isomorphisms (see section 10). It reduces to the usual group if M is simply connected.

It will be seen in the proofs that the problem we are considering is a generalization of the problem of extending continuous mappings. We shall both generalize and use extensively certain theorems of Eilenberg in this connection [2].

In an appendix, the existence and properties of the characteristic cohomology class are established for the more general type of fibre space introduced by Hurewicz and the author [3]. Any lengthy proof that would interrupt the trend of ideas is postponed to a later section.

¹ Numbers in square brackets refer to the bibliography.

² Whitney [7] has proved this by showing that any C^r -manifold is C^r -homeomorphic to an analytic manifold in a euclidean space.

2. Spaces of point tensors

In the following, M will denote the differentiable manifold³ of dimension n over which tensors are to be defined. The class r of M is assumed to be ≥ 2 . It is not assumed that M is compact. It is assumed that M is separable so that a countable set of coördinate neighborhoods can be found to cover it. It follows from results of Cairns [1] that M can be triangulated. We may therefore assume that M is homeomorphic with a definite simplicial countable, complex. The symbol M^q will denote the subcomplex of M composed of its simplexes of dimensions $\leq q$.

In order to be explicit and yet not have too cumbersome a notation, we shall define only the manifold M' of point tensors over M having one contravariant and two covariant indices. If P is a point of M , $C(x)$ a coördinate neighborhood of P , and a_{jk}^i ($i, j, k = 1, \dots, n$) an ordered set of n^3 real numbers, the combination $(P, C(x), a_{jk}^i)$ is called a *representation of a point tensor*. Two such $(P, C(x), a_{jk}^i), (Q, C(\bar{x}), \bar{a}_{jk}^i)$ are said to be equivalent if $P = Q$ and

$$(A) \quad \bar{a}_{jk}^i = a_{vw}^u \frac{\partial x^v}{\partial \bar{x}^j} \frac{\partial x^w}{\partial \bar{x}^k} \frac{\partial \bar{x}^i}{\partial x^u}.$$

(The derivatives are evaluated at P .) As is well known, this is a proper equivalence relation, so that the combinations $(P, C(x), a_{jk}^i)$ fall into mutually exclusive equivalence classes. An equivalence class is called a *point tensor*. The adjective *point* is used for the obvious reason that a point tensor is attached to that point of M occurring in any one of its representations.

The family of all point tensors of the above type at all points of M form a set M' which we now proceed to show is a differentiable manifold in a natural way. To do this we must define 1-1 maps of subsets of M' into suitable subsets of a euclidean space. Let $C(x)$ be an admissible coördinate neighborhood in M . Let U be the family of all point tensors having representations $(P, C(x), a_{jk}^i)$ (where P varies in $C(x)$, and the a 's are arbitrary). Then U is in 1-1 correspondence with this set of representations in $C(x)$. Attach to $(P, C(x), a_{jk}^i)$ the $n + n^3$ real numbers $x^1, \dots, x^n, a_{jk}^i$ where the x 's are the coördinates of P in $C(x)$. The two correspondences define a 1-1 map of U into the open subset of $(n + n^3)$ -euclidean space of coördinates (x, a) where x is in $C(x)$ and the a 's are arbitrary. Thus for each $C(x)$ in M a coördinate neighborhood $C(x, a)$ in M' is determined. If a point tensor is in both $C(x, a)$ and $C(\bar{x}, \bar{a})$ then, if $\bar{x}^i = f^i(x)$ is the coördinate transformation from $C(x)$ to $C(\bar{x})$, these equations together with the equations (A) define the coördinate transformation from $C(x, a)$ to $C(\bar{x}, \bar{a})$. Thus, the \bar{x} 's are functions of the x 's alone while the \bar{a} 's are functions of both the x 's and the a 's and are linear in the latter. The determinant of this transformation is easily seen to be a power of the determinant of the transformation from $C(x)$ to $C(\bar{x})$. Since the derivatives in (A) are of class

³ For the concept of differentiable manifold see [7].

$r - 1$, it follows that M' is a differentiable manifold of class $r - 1$. Its dimension is $n + n^3$.

It is clear that the relative point tensors of a given weight likewise form a differentiable manifold. The same procedure likewise carries through for the space of point affine connections. It is only necessary that the transformation law in question shall define a proper equivalence relation among the representations. For example, ordered pairs of covariant vectors at points of M have representations of the form $(P, C(x), a_i, b_i)$ with coördinate transformations of the form

$$\bar{x}^i = f^i(x), \quad \bar{a}_i = a_u \frac{\partial x^u}{\partial \bar{x}^i}, \quad \bar{b}_i = b_u \frac{\partial x^u}{\partial \bar{x}^i}.$$

The determinant is again a power of the determinant of $x \rightarrow \bar{x}$.

In the following pages we shall frequently speak of "a tensor manifold M' over M ". It is to be understood that M' is the space of point tensors over M of a fixed order and weight—or M' is the space of point affine connections over M —or M' is the space of ordered sets of vectors (or tensors of given order and weight) at points of M .

3. The projection of the tensor manifold onto M

Let the function π assign to each point tensor in M' the point of M to which the tensor is attached. This natural mapping of M' onto M we refer to as the *projection*. If $C(x)$ is a coördinate neighborhood in M , and $C(x, a)$ the corresponding one in M' , then π has the form of a projection in these coördinates. The Jacobian of the projection has rank n at each point of M .

The inverse image of a point P of M in M' (i.e. the point tensors at P) form a linear space which we shall refer to as the fibre F_P over P . It follows readily that M' is a fibre bundle over M in the sense of Whitney [6]. The fibres are euclidean spaces with linear transformations as the admissible group, and the fibres over a particular coördinate neighborhood form a product space.

4. Tensor functions over M

A *tensor function* over M is a map f of M in a tensor manifold M' over M having the property that πf is the identity. The image of M in M' under f is called the *graph* of f .

The usual definition of tensor function has the disadvantage that graphs exist only locally and then vary with the coördinate system. The advantage of the present method is that, by identifying the equivalent functional values first, the function itself falls into that class of objects (usually referred to as functions) which are point to point correspondences.

If f is a map of the set A in the set B , the graph of f is usually defined to be the set $(a, f(a))$ in the product space $A \times B$. Therefore the foregoing definition of graph of a tensor function needs a few words of justification. Suppose M' is the space of point scalars over M . It is not hard to see that M' is the product

space $M \times L$ of M with a coordinate line L . Any map f of M in L (i.e. a scalar function in the usual sense) has a graph in $M \times L$ which defines a scalar function in the sense described above. Therefore, at least in this case, the term is justified. An additional reason is that for any neighborhood $C(x)$ the part of the graph over it is the graph of the functions $a(x)$ in the product space $C(x, a)$.

In general, as examples given later will show, M' is not the product space of M with a fibre. Consequently we cannot expect that tensor functions as usually defined will have graphs in the usual sense. We have, however, with the present definition, most of the important features of a graph. The graph is in a space which is locally a product space over M , and we have the projection π which, when applied to the graph, results in the identity map of M . The only missing feature is the resolution of M' into a product space of a second space with M . This, however, is inherent in the nature of tensor.

In case M can be resolved into a product space, it is of little interest unless the resolution satisfies several strong conditions. We shall say that M' is *properly resolved* into a product space $M \times E$ where E is a euclidean space provided there is a homeomorphic map of class $r - 1$ of $M \times E$ onto M' with the property that, for each $P \in M$, the section $P \times E$ is mapped linearly on the fibre over P in M . Then we have

THEOREM 1. *If M' is a tensor space over M of dimension $n + m$, it can be properly resolved into a product space if and only if there exist m tensor functions over M of class $r - 1$ which are linearly independent at each point of M .*

Suppose M is properly resolved, and ϕ is its homeomorphism with $M \times E$. Let a_1, \dots, a_m be m linearly independent points of E . Then $f_i(P) = \phi(P, a_i)$ ($i = 1, \dots, m$) are m tensor functions of class $r - 1$ which are linearly independent at each point of M .

Conversely, let $\{f_i(P)\}$ be m independent tensor functions of class $r - 1$. Define $\phi(P, a_1, \dots, a_m) = \sum a_i f_i(P)$. Then ϕ is a proper map of $M \times E$ onto M .

5. Statement of problem

The most general type of problem for which we shall give a method of attack may be formulated as follows:

Let M' be a tensor manifold over M , and let M'' be a differentiable submanifold of M' such that the projection π maps M'' onto M , has a Jacobian of rank n at every point of M'' , and M'' is locally a product space over M . (This last means that each point P of M has a neighborhood C such that the product $C \times F''$, where F'' is the fibre over P , is homeomorphic with the part of M'' over C , and, for any point Q of C , $Q \times C$ is mapped on the fibre over Q .) The problem is to define a map f of M in M'' such that πf is the identity map of M , and f has a specified class (up to the class of M'').

The following are examples of problems of this type.

a. If M'' is all of M' , the solution is immediate, for the zero tensor at every point is an f of the required kind.

b. Let M' be the space of contravariant point vectors over M and M'' the family of non-zero vectors. This is the problem of defining a non-singular vector field over M .

c. Let M' be the covariant point tensors of order 2, and M'' the symmetric positive definite ones. A solution f determines a Riemann metric in M .

d. Let M' be as in c, and let M'' be the symmetric semi-definite tensors of a fixed rank k .

e. Let M' be as in c, and let M'' be the symmetric non-singular tensors of a fixed index h .

f. Let M' be the set of ordered sets of k contravariant vectors, ($k \leq n$), and let M'' be the subspace of independent sets. This is the problem considered by Stiefel [5].

g. Consider the problem raised by Theorem 1. Let M'' be the tensor space of ordered sets of m tensors each of the same order as in M' . Let M''' be the subspace of independent sets. Then the existence of a map f of M into M''' such that $\pi f = \text{identity}$ is a necessary and sufficient condition that M' can be properly resolved into a product space.

The methods to be given seem to be of little value in handling a problem involving differential conditions such as defining a flat affine connection or a Riemann metric of constant or constant mean curvature.

The following theorem reduces the general problem to the simpler one of finding a continuous admissible tensor defined on M .

THEOREM 2. *If f is a continuous map of M in M'' such that $\pi f = \text{identity}$, then there exists a map f' of class ∞ such that $\pi f' = \text{identity}$. If a metric is given in M'' , f' may be chosen to approximate f as closely as desired. (If M'' is analytic, we can only assert the existence of an f' of class ∞).*

6. Construction of a Riemann metric

In order to illustrate the basic method of attack on our problem, we shall show that one can define over M a symmetric, positive definite, tensor function g of covariant order 2. This result has already been established by Whitney [7]. His procedure is to construct a regular imbedding of M in a euclidean space and then to take over into M the metric of the euclidean space. Our procedure is a direct construction of the metric tensor and is somewhat simpler. However, it fails to give the imbedding theorem.

Let M' be the manifold of covariant point tensors of order 2 over M . Let M'' be the subset of symmetric, positive definite ones. Let F' be the fibre in M' over a point P in M , and let F'' be its part in M'' . The important observation at this point is that F'' is a convex linear cell of dimension $n(n+1)/2$. This is proved by noting that a linear combination with positive coefficients of two positive definite quadratic forms is positive definite. Since a symmetric matrix is determined by $n(n+1)/2$ of its elements, and since any symmetric matrix sufficiently near a positive definite one is likewise positive definite, it follows that the dimension is $n(n+1)/2$.

We suppose now that M is subdivided into a simplicial complex so fine that each simplex is contained wholly within some one coördinate neighborhood. Since the number of simplexes is countable, we may order them in a sequence $\sigma_1, \sigma_2, \dots$ in such a way that a simplex is preceded by any one of its faces. Then σ_1 is a vertex P of M . We choose a point in the fibre F'' over P and denote it by $g(P)$. Suppose inductively that g has been defined over all simplexes preceding σ_k so as to be continuous on their point set sum and $\pi g = \text{identity}$. Then g is defined on $\partial\sigma_k$ (the boundary of σ_k). Let $C(x)$ be a coördinate neighborhood containing σ_k . Then the functions $g_{i,j}(x)$ are defined and continuous over $\partial\sigma_k$. These functions alone map $\partial\sigma_k$ into the fibre F'' over a point P in $C(x)$. Since F'' is a cell, these functions can be extended over the interior of σ_k so as to give a continuous map of the closed simplex σ_k in F'' . The desired map g of σ_k in $C(x, g_{i,j})$ is now obtained by setting $g(x) = (x, g_{i,j}(x))$ for $x \in \sigma_k$. Then g is continuous on the sum of the first k simplexes, for if a function is continuous on each of two compact sets, it is continuous on their sum. This inductive construction leads to a continuous map g of M in M'' . The differentiable approximation is given by Theorem 2.

7. Observations on the construction

It is clear that the construction just given applies in any problem where F'' is a cell. This condition may be weakened to the extent of requiring that every continuous map of a $(k-1)$ -sphere in F'' for $k \leq n$ is contractible to a point in F'' . This latter condition is equivalent to the requirement that all homotopy groups of F'' of dimensions $< n$ shall vanish. This in turn is equivalent to the same requirement on the homology groups of F'' of dimensions $< n$ [4].

As an application, suppose it is required to define a tensor function over M of a prescribed order > 1 which is not zero at any point. In this case the fibre F'' is a euclidean space with its origin deleted. Since the order is > 1 , the dimension m of the euclidean space is $> n$. Deleting a single point from a euclidean m -space does not alter the fact that a map of a sphere of dimension $< m-1$ in the space is contractible to a point. The required function therefore exists.

We formulate our results as follows:

THEOREM 3. *If all the homotopy groups or equally well all the homology groups of the fibre F'' vanish for each dimension $< n$, then there exists a tensor defined over M of the specified type.*

8. Construction of the characteristic cocycle

If some sphere in F'' of dimension $< n$ is not contractible, we can expect essential difficulties. In order to discuss this case we make the

ASSUMPTIONS. Let h be the smallest integer such that the homotopy group $[4], \pi_h(F'') \neq 0$. We suppose $h < n$. By $\pi_0(F'')$ is meant the 0th homology group of F'' where cycles are taken with integer coefficients, and the sum of the coefficients of a cycle is zero. If $h = 1$, we shall suppose that $\pi_1(F'')$ is abelian.

Since M'' is locally a product space, any two nearby fibres are homeomorphic. As M is connected we can pass from one fibre to any other through a succession of fibres such that successive pairs are homeomorphic. Thus, the above assumptions need only have been made of one fibre.

It is a consequence of our assumptions that any map of an oriented h -sphere in F'' determines a unique element of $\pi_h(F'')$ (i.e. it is not necessary that a fixed reference point of the sphere be mapped on a fixed reference point of F'').

We shall suppose that M is subdivided so fine that the star of a simplex lies wholly within a single admissible neighborhood of M (one such that the part of M'' over it is a product $C \times F''$). Then, by the argument of section 7, one may construct a map f of the h -dimensional skeleton M^h in M'' such that $\pi f = \text{identity}$.

To any map f of M^h in M'' such that $\pi f = \text{identity}$, we attach a chain $c^{h+1}(f)$ of M as follows. If σ is an oriented $(h+1)$ -simplex, C an admissible neighborhood containing σ , and $C \times F''$ a representation of $\pi^{-1}(C)$ in M'' , then f maps $\partial\sigma$ into $C \times F''$. Let λ map $C \times F''$ into F'' by attaching to each point its F'' coördinate. Then λf maps $\partial\sigma$ into F'' , determining thereby an element of $\pi_h(F'')$ which we denote by $c(f, \sigma)$. Then⁴

$$c^{h+1}(f) = \sum c(f, \sigma^{h+1}) \sigma^{h+1}.$$

The sum extends over all $(h+1)$ -simplexes of M .

An object is a *chain* if it is a function from simplexes to a group. Here F'' varies from one neighborhood to another, so the definition of chain is not satisfied by $c^{h+1}(f)$. The definition will be satisfied when we have established a fixed reference F''_0 , and fixed isomorphisms connecting $\pi_h(F'')$ for each F'' to $\pi_h(F''_0)$. We know such isomorphisms exist; but arbitrarily chosen ones will not serve our purpose. We consider this matter in the following two sections.

9. The orientable case

Let A be a curve joining two points P, Q of M defined by a function $\phi(t)$, $a \leq t \leq b$, $\phi(a) = P$, $\phi(b) = Q$. Suppose a continuous function $h(y, t)$ is given with values in M'' , defined for y in F''_P and $a \leq t \leq b$, such that $\pi h(y, t) = \phi(t)$ and $h(y, 0) = y$. Then we shall say that h deforms F''_P along A into F''_Q . Two such deformations of F''_P are said to be *equivalent* if the two resulting maps of F''_P in F''_Q are homotopic in F''_Q .

If the curve A lies in a neighborhood C and the part of M'' over C is represented as a product $C \times F''$, then $h(y, t)$ can be defined as the point $(\phi(t), y)$ of the product $C \times F''$. Now any curve A can be expressed as a sum of curves A_1, \dots, A_k each of which lies in a single neighborhood. By piecing together homotopies along each, one obtains a homotopy along A .

A homotopy of F''_P along A into F''_Q induces an isomorphism between $\pi_h(F''_P)$

⁴ The chain $c^{h+1}(f)$ is the "characteristic cocycle" to be found in the work of Stiefel [5] and Whitney [6].

and $\pi_h(F''_Q)$. This isomorphism appears to depend on several factors. However, we have

LEMMA 1. *If A_1, A_2 are two curves from P to Q which are homotopic leaving their end points fixed, then a deformation of F''_P along A_1 into F''_Q is equivalent to one along A_2 .*

The virtue of the lemma is that it assures us that the isomorphism set up between $\pi_h(F''_P)$ and $\pi_h(F''_Q)$ by deforming F''_P along a curve into F''_Q depends only on the homotopy class of the curve. Consider now the case $P = Q$. Then every element of $\pi_1(M)$ defines an automorphism of $\pi_h(F''_P)$. It is not hard to see that the attached automorphism gives a homomorphism of $\pi_1(M)$ into the group of automorphisms of $\pi_h(F''_P)$.

DEFINITION. We say that M'' is *orientable* over M relative to $\pi_h(F'')$ if, for each point P of M and each element of $\pi_1(M)$, the resulting automorphism of $\pi_h(F''_P)$ is the identity.

The use of the word *orientable* is explained by the fact that M itself is orientable if and only if the space M'' of non-zero, contravariant, point vectors is orientable over M relative to $\pi_{n-1}(F'')$.

The next remark is that in the definition of orientability we do not need to require of each point P that any element of $\pi_1(M)$ shall induce the identity automorphism. It is sufficient to demand this of just one point P_0 ; it then follows for every point P . This follows from the well-known fact that any closed curve beginning and ending at P is homotopic to a path which first describes a curve A from P to P_0 , then a closed curve from P_0 to P_0 , then finally describes A^{-1} .

The virtue of the orientable case is that any two deformations of F''_P along curves into F''_Q induce the same isomorphism between $\pi_h(F''_P)$ and $\pi_h(F''_Q)$. For, if A, B are two curves from P to Q , and ϕ_1, ϕ_2 are the two corresponding isomorphisms, then deforming F''_Q around the closed curve $A^{-1}B$ gives the identity isomorphism of $\pi_h(F''_Q)$. This automorphism applied to ϕ_1 gives ϕ_2 , since the path $AA^{-1}B$ is homotopic to B .

Assuming now the orientable case, we choose a fixed fibre F''_0 , and for any fibre F'' we set up the unique isomorphism between $\pi_h(F'')$ and $\pi_h(F''_0)$ obtained by deforming F'' along a curve into F''_0 . In this way, each element of $\pi_h(F'')$ for any F'' represents a unique element of $\pi_h(F''_0)$. Consequently, in the orientable case, the object $c^{k+1}(f)$ defined above is a chain (in the proper sense of the word) with coefficients in the group $\pi_h(F''_0)$.

10. The non-orientable case

Here we cannot proceed as before since no unique isomorphisms in the large exist. However, for each simplex σ , unique isomorphisms connecting the $\pi_h(F''_P)$ for $P \in \sigma$ can be set up using curves in σ ; for a simplex is simply connected. Thus each simplex σ of M has a coefficient group $G(\sigma)$. A q -chain is now defined to be a function f attaching to each q -simplex σ an element $f(\sigma) \in G(\sigma)$. The q -chains form a group in the obvious way. Suppose σ' is a face of σ ($\sigma' < \sigma$),

then by using a curve in the closure of σ , joining a point of σ to a point of σ' , a unique isomorphism $h_{\sigma'\sigma}$ of $G(\sigma)$ onto $G(\sigma')$ can be defined. If $[\sigma:\sigma']$ is the incidence number, we can now define the boundary of a q -chain f to be the $(q-1)$ -chain ∂f having on σ' the value

$$\partial f(\sigma') = \sum_{\sigma} [\sigma:\sigma'] h_{\sigma'\sigma}(f(\sigma))$$

which is an element of $G(\sigma')$. Since the closure of σ is simply connected, $\sigma'' < \sigma' < \sigma$ implies $h_{\sigma''\sigma'} h_{\sigma'\sigma} = h_{\sigma''\sigma}$. From this it follows that $\partial \partial f = 0$. Co-boundary is defined by

$$\delta f(\sigma) = \sum_{\sigma'} [\sigma:\sigma'] h_{\sigma'\sigma}^{-1}(f(\sigma'));$$

and, again, $\delta \delta f = 0$. Thus cycles, cocycles, homology and cohomology can be defined as usual.⁵ Thus, in the non-orientable case, we are dealing with the local coefficient groups $\pi_h(F_P'')$ connected by the local isomorphisms gotten by deforming the F_P'' along curves.⁶

11. Properties of the characteristic cocycle

Having agreed on the sense in which $c^{h+1}(f)$ can be regarded as a chain in M , we are prepared to state its principal properties.

THEOREM 4. Under the assumptions of section 8, there exist maps f of the h -dimensional skeleton M^h of M in M'' such that $\pi f = \text{identity}$. Any such f defines a chain $c^{h+1}(f)$ with the following properties:

(a). $c^{h+1}(f) = 0$ is a necessary and sufficient condition that f can be extended continuously to M^{h+1} , preserving $\pi f = \text{identity}$.

(b). $c^{h+1}(f)$ is a cocycle.

(c). If f' is any other map of M^h in M'' such that $\pi f' = \text{identity}$, then $c^{h+1}(f') \smile c^{h+1}(f)$.

(d). If c^{h+1} is a cocycle cohomologous to $c^{h+1}(f)$, there exists a map f' of M^h in M'' such that $\pi f' = \text{identity}$, and $c^{h+1} = c^{h+1}(f')$.

(e). A necessary and sufficient condition that there be a map f of M^{h+1} in M'' such that $\pi f = \text{identity}$ is that, for any such map f' of M^h , we have $c^{h+1}(f') \smile 0$.

⁵ The resulting groups may be handled much as the ordinary ones. Care must be taken that simplicial transformations are accompanied by local coefficient group isomorphisms. The usual proof of topological invariance carries through with just these modifications. These new groups may also be defined for a general space. Here each point P has its coefficient group $G(P)$, and a homotopy class of curves from P to Q determine an isomorphism which is transitive under addition. It has come to the author's attention that "local coefficients" and Reidemeister's "Überdeckungen" (*Topologie der Polyeder*, Leipzig, 1938) are equivalent concepts. A discussion of the connection and an analysis of this homology theory will be found in a forthcoming paper.

⁶ The procedure of Whitney [6], in the case of non-orientable sphere-bundles, is to reverse the sign of certain incidence numbers in M . His method can apply in our case only when the sole automorphism on $G(\sigma)$ is a reversal of sign.

(f). The cohomology class of $c^{h+1}(f)$ is independent of the subdivision of M which is used and of the map f of M^h ; it is therefore a topological invariant of the pair of spaces M, M'' and the map π .

12. Proof^{6a} of Theorem 2

In order to construct the differentiable approximation, we shall prove some preliminary results.

(A). Let D, D' be two rectangular domains in n -space E^n defined by $a_i < x_i < b_i, a'_i < x_i < b'_i$, respectively, and such that D' contains the closure \bar{D} of D . Then there exists a real-valued function g defined in E^n of class ∞ , and such that

$$0 \leq g \leq 1, \quad g = \begin{cases} 1 & \text{in } D, \\ 0 & \text{in } E^n - D'. \end{cases}$$

Let (c, d) be an interval and let

$$\psi_{cd}(x) = \begin{cases} \exp\left(-\frac{1}{x-c} + \frac{1}{x-d}\right) & \text{in } (c, d), \\ 0 & \text{outside } (c, d). \end{cases}$$

Then ψ is of class ∞ , and $\psi \geq 0$. Let

$$\phi_{cd}(x) = \int_c^x \psi_{cd}(t) dt / \int_c^d \psi_{cd}(t) dt.$$

Then ϕ is of class ∞ , $0 \leq \phi \leq 1$, $\phi = 0$ for $x \leq c$, and $\phi = 1$ for $x \geq d$. If $(a, b), (a', b')$ are two intervals and $a' < a, b < b'$, then, by piecing together two such functions as ϕ , we obtain a function

$$g(x) = \begin{cases} \phi_{a'a}(x) & \text{for } x \leq b, \\ 1 - \phi_{bb'}(x) & \text{for } x > b \end{cases}$$

of class ∞ , $0 \leq g \leq 1$, $g = 1$ in (a, b) , $g = 0$ outside (a', b') . Let $g_i(x_i)$ be such a function for the pair $(a_i, b_i), (a'_i, b'_i)$. Then the product $g(x_1, \dots, x_n) = g_1(x_1) \cdots g_n(x_n)$ has the properties asserted in (A).

(B). Let U be an open set in E^n with compact closure \bar{U} , and let U' be an open set containing \bar{U} . Then there exists a real-valued function g defined in E^n of class ∞ , and such that

$$0 \leq g \leq 1, \quad g = \begin{cases} 1 & \text{on } \bar{U}, \\ 0 & \text{outside } U'. \end{cases}$$

As \bar{U} is compact, we can choose a finite number D_1, \dots, D_m of rectangular domains covering \bar{U} such that the closure of each is in U' . Let D'_i be a rec-

^{6a} A shorter proof can be based on a result of H. Whitney, *Analytic extensions of differentiable functions defined in closed sets*, Trans. Amer. Math. Soc., 36 (1934), pp. 63-89, Th. 3.

tangular domain containing \bar{D}_i and contained in U' . Then there is a function g_i attached to the pair D_i, D'_i as in (A). Define the function g in E^n by

$$1 - g = (1 - g_1)(1 - g_2) \cdots (1 - g_m).$$

Then g is of class ∞ , $0 \leq g \leq 1$, some $g_i = 1$ implies $g = 1$, every $g_i = 0$ implies $g = 0$. Thus $g = 1$ in $\sum D_i$, and $g = 0$ outside $\sum D'_i$.

(C). If Q is a point of M'' , and $C(x)$ a coordinate neighborhood of $\pi(Q)$, then a coordinate neighborhood $C''(y)$ of Q in M'' may be chosen such that π has the form $x^i = \pi^i(y) \equiv y^i$ ($i = 1, \dots, n$).

Let $C(\bar{y})$ be a neighborhood of Q , and let π be given by the functions $x^i = \bar{\pi}^i(\bar{y})$. Since π has rank n at Q , the equations $y^i = \bar{\pi}^i(\bar{y})$ can be solved for n of the \bar{y} 's, say $\bar{y}_1, \dots, \bar{y}_n$, as functions of the y 's and the other \bar{y} 's in some neighborhood of Q . If we let $y^i = \bar{y}^i$ ($i = n + 1, \dots, m$), then, in this neighborhood, the y 's form an admissible coordinate system.

Returning now to the proof proper of Theorem 2, let f be a map of M in M'' such that $\pi f = \text{identity}$. Let $P \in M$ and $f(P) = Q \in M''$. Corresponding to a neighborhood C of P , let C'' be the neighborhood of Q given by (C). We can then choose two neighborhoods D, E of P such that $C \supset \bar{D}$, $D \supset \bar{E}$, and $f(\bar{D}) \subset C''$. As M is locally compact and separable, we may choose a sequence $\{C_i, C'_i, D_i, E_i\}$ of such sets of neighborhoods such that M is covered by the neighborhoods E_i , and each \bar{E}_i meets only a finite number of other \bar{E}_j . This latter property can also be arranged for the D_i by reducing them in size without losing the property $D_i \supset \bar{E}_i$.

In the pair of neighborhoods D_1, C'_1 , f is given by m functions $y^i = f^i(x)$. Due to the choice of coordinates in C'_1 and $\pi f = \text{identity}$, we have $f^i(x) \equiv x^i$ ($i = 1, \dots, n$). Choose a pair of neighborhoods U', U such that $D_1 \supset \bar{U}'$, $U' \supset \bar{U}$ and $U \supset \bar{E}_1$. If $\epsilon > 0$ is given, we can choose functions $\phi^j(x)$ ($j = n + 1, \dots, m$) in D_1 of class ∞ and such that $|\phi^j - f^j| < \epsilon$. Let $g(x)$ be the function given by (B) for the pair U', U . Define

$$\psi^j = g\phi^j + (1 - g)f^j \quad \text{in } D_1.$$

Then ψ^j is of class ∞ in U , and $\psi^j = f^j$ outside U' . If ϵ is sufficiently small, the functions $f^1, \dots, f^n, \psi^{n+1}, \dots, \psi^m$ define a map f_1 of D_1 in C'_1 . If $f_1 = f$ outside D_1 , then f_1 is a map of M in M'' such that $\pi f_1 = \text{identity}$, and f_1 is differentiable to the class of M'' in an open set about \bar{E}_1 . Since only a finite number of \bar{D}_i meet \bar{D}_1 , by restricting ϵ , we can insure that f_1 maps \bar{D}_i in C'_i for each i .

Suppose a map f_k of M in M'' is given such that $\pi f_k = \text{identity}$, f_k maps each \bar{D}_i in C'_i , and f_k is differentiable to the class of M'' in an open set W about $\sum_{i=1}^k \bar{E}_i$. Choose a pair U', U of open sets such that $U \supset \bar{E}_{k+1} - W \cdot \bar{E}_{k+1}$, $U' \supset \bar{U}$, $D_{k+1} \supset \bar{U}'$ and $\bar{U}' \cdot \sum_{i=1}^k \bar{E}_i = 0$. Apply now the construction of the preceding paragraph to each of the last $m - n$ components of f_k in the neighborhood D_{k+1} . The components ψ_k^j will not only be differentiable in U but likewise wherever the f_k^j are differentiable, namely, in $W \cdot D_{k+1}$. This leads to a new map f_{k+1} which is differentiable in an open set about $\sum_{i=1}^{k+1} \bar{E}_i$.

Since $f_{k+1} = f_k$ on $\sum_1^k \bar{E}_i$, the sequence of functions $\{f_k\}$ so constructed converges to a map f' of M in M'' such that $f' = f_k$ on $\sum_1^k \bar{E}_i$. Therefore f' is differentiable to the class of M'' everywhere and $\pi f' = \text{identity}$.

13. Proof of Lemma 1

Consider the closed curve $A_1^{-1}A_2$ beginning and ending at Q . By assumption it is contractible into Q leaving its end points fixed. The two maps of F''_P in F''_Q are homotopic along this curve. We must cover the contraction of $A_1^{-1}A_2$ into Q by a contraction of the homotopy into F''_Q . Let E be a 2-cell, and ψ a map of E in M so that its boundary ∂E is mapped into $A_1^{-1}A_2$. Subdivide E into simplexes so fine that the image of each lies wholly within a coordinate neighborhood of M . Now $A_1^{-1}A_2$ can be contracted into Q over $\psi(E)$ in a finite number of steps, each consisting of deforming an arc of the curve over a simplex of E . Thus, each step occurs in a single coordinate neighborhood. Suppose then $\phi(t, \theta)$, ($a \leq t \leq b$, $0 \leq \theta \leq 1$) is a homotopy of the arc $\phi_0(t)$ into $\phi_1(t)$ leaving the end points fixed and occurring in a neighborhood $C(x)$. Let $h(y, t)$ be a homotopy of F''_P along $\phi_0(t)$. Now the part of M'' over $C(x)$ is a product space $C \times F''$. Hence, $h(y, t)$ is given by a pair of functions $h'(y, t) = \phi_0(t)$ in C and $h''(y, t)$ in F'' . Define $h'(y, t, \theta) = \phi(t, \theta)$ and $h''(y, t, \theta) = h''(y, t)$. The second pair h', h'' define in $C \times F''$ a deformation of $h(y, t)$ into a homotopy $h_1(y, t)$ along $\phi_1(t)$ without altering $h(y, a)$ and $h(y, b)$. This proves the lemma.

14. Proof of Theorem 4

The proofs of (a) to (e) parallel closely proofs to be found in a paper by Eilenberg [2]. The modifications are all of the same type. We shall be dealing with a simplex σ of M ; C will be a neighborhood of σ , $C \times F''$ will be a representation of the part of M'' over C . In each case our problem will be to construct, extend, or deform homotopically a map f of σ or $\partial\sigma$. The function f has, locally, two components $\pi f = \text{identity}$ in C , and λf in F'' . It will be seen that Eilenberg's arguments apply in each case to the component λf .

(a). Suppose f is defined on M^{h+1} . Then for any σ^{h+1} , $\lambda f(\sigma^{h+1})$ is a cell in F'' whose boundary is $\lambda f(\partial\sigma^{h+1})$. Therefore $c(f, \sigma^{h+1}) = 0$. Conversely, $c(f, \sigma^{h+1}) = 0$ means that λf can be extended to a map of σ^{h+1} in F'' . If we extend πf over σ^{h+1} to be the identity, we obtain the extension of f .

(b). Let σ^{h+2} be arbitrary, and $K = |\sigma^{h+2}|$. Now λf maps K^h in F'' . If F'' is taken to be the space Y of Eilenberg, then the chain $c^{h+1}(\lambda f)$ of Eilenberg is identical with $c^{h+1}(f)$ over K . As the former is known to be a cocycle on K , it follows that σ^{h+2} enters $\delta c^{h+1}(f)$ with coefficient zero. As σ^{h+2} is arbitrary, the proof is complete.

(c). We need the following:

LEMMA 2. *There exists a homotopy $f(x, t)$ of $f(x)$ for $x \in M^h$, $0 \leq t \leq 1$, such that $\pi f(x, t) = x$ for $x \in M^h$ and all t , $f(x, 0) = f(x)$ for $x \in M^h$, and $f(x, 1) = f'(x)$ for $x \in M^{h-1}$.*

Define $f(x, 0) = f(x)$ for $x \in M^h$, $f(x, 1) = f'(x)$ for $x \in M^{h-1}$ and $\pi f(x, t) = x$ for $x \in M^h$, $0 \leq t \leq 1$. We must extend the local components $\lambda f(x, t)$ over the rest of $M^h \times I$. Order the simplexes of M^{h-1} in a sequence such that each simplex is preceded by its faces. As the first is a vertex P , and F'' is connected (if $h > 0$), we can join $f(P)$ to $f'(P)$ by a curve $f(P, t)$ in F'' . Suppose $f(x, t)$ is defined over all prisms $\sigma \times I$ for simplexes σ preceding σ_i . Then $f(x, t)$ is defined on $\sigma_i \times 0$, $\sigma_i \times 1$ and $\partial \sigma_i \times I$, and therefore $\lambda f(x, t)$ maps $\partial(\sigma_i \times I)$ in F'' . As the dimension of this sphere is $< h$, λf can be extended over $\sigma_i \times I$. Having thus defined $f(x, t)$ over $M^{h-1} \times I$, we have only to extend it over each $\sigma^h \times I$.

As $\lambda f(x, t)$ is defined over $\sigma^h \times 0 + \partial \sigma^h \times I$, and this set is a retract of $\sigma^h \times I$, $\lambda f(x, t)$ admits a continuous extension over $\sigma^h \times I$. This proves the lemma.

Let $f''(x) = f(x, 1)$. Since for any σ^{h+1} , $\lambda f(\partial \sigma^{h+1})$ is homotopic in F'' to $\lambda f''(\partial \sigma^{h+1})$, we have $c^{h+1}(f) = c^{h+1}(f'')$. Corresponding to the maps f' , f'' which agree on M^{h-1} , we define (following Eilenberg) an h -chain $d^h(f', f'')$. Let σ^h be any h -simplex. Write the oriented h -sphere as a sum $S^h = E' + E''$ of two hemispheres. Let g' , g'' map E' , E'' respectively on σ^h topologically with degrees 1, -1 so that they agree on the common boundary. Then define $\phi = f'g'$ on E' , $\phi = f''g''$ on E'' . As f' , f'' agree on $\partial \sigma^h$, $\lambda \phi$ is a continuous map of S^h in F'' determining thereby an element of $\pi_h(F'')$ which is denoted by $d(f', f'', \sigma^h)$. Let $d^h(f', f'') = \sum d(f', f'', \sigma^h) \sigma^h$. Then

$$(A) \quad \delta d^h(f', f'') = c^{h+1}(f') - c^{h+1}(f'').$$

If, for any σ^{h+1} , we consider the functions $\lambda f'$, $\lambda f''$ defined on $|\partial \sigma^{h+1}|$ and the chains $d^h(\lambda f', \lambda f'')$, $c^{h+1}(\lambda f')$, $c^{h+1}(\lambda f'')$ as defined by Eilenberg for the complex $|\sigma^{h+1}|$, then the relation analogous to (A) holds. However these latter chains agree with the former where both are defined. The proof of (c) is therefore complete.

(d). Let $\delta d^h = c^{h+1} - c^{h+1}(f)$. Define $f' = f$ on M^{h-1} . Given any σ^h , define $\pi f' = \text{identity}$ in σ^h . Then define $\lambda f'$ on σ^h so that $d(f', f, \sigma^h)$ (see proof of (c)) is the coefficient of σ^h in d^h . Then $d^h(f', f) = d^h$. As shown in (c), $\delta d^h(f', f) = c^{h+1}(f') - c^{h+1}(f)$. Therefore, $c^{h+1} = c^{h+1}(f')$.

(e). If a map f of M^{h+1} exists, then, by (a), $c^{h+1}(f) = 0$. By (c), for any f' defined on M^h , $c^{h+1}(f') \cup c^{h+1}(f) = 0$. Suppose $c^{h+1}(f') \cup 0$ for some map of M . Then, if we choose $c^{h+1} = 0$ in (d), we obtain a new map f'' of M^h such that $c^{h+1}(f'') = 0$. It follows from (a) that f'' can be extended over M^{h+1} .

(f). Let M_1 , M_2 be two subdivisions of M into complexes. We suppose M_2 is so fine that there is a simplicial map τ of M_2 into M_1 such that, for each point x , $\tau(x)$ lies on the closure of the simplex of M containing x . There is, therefore, a homotopy $\tau(x, t)$ connecting the identity to $\tau(x)$ such that $\tau(x, t)$ lies on the closure of the simplex of M_1 containing x for $0 \leq t \leq 1$. We now suppose M_1 is so fine that the map τ of any prism $\zeta \times I$, where ζ is a simplex of M_2 , lies in a neighborhood of M . (This will be the case if the closure of the star of each simplex of M_1 lies in a neighborhood). Now let f be a map of M_1^h in M'' . Construct a map f' of M_2^{h-1} in M'' (always so that $\pi f' = \text{identity}$). We must extend

f' over M_2^h so that $c_2^{h+1}(f') \smile c_1^{h+1}(f)$. To this end we shall define a map $f'(x, t)$ of $M_2^h \times I$ in M'' . Let $f'(x, 1) = f\tau(x)$ for $x \in M_2^h$, and $f'(x, 0) = f'(x)$ for $x \in M_2^{h-1}$. Order the simplexes of M_2^h in a sequence so that each is preceded by its faces. If P is the first (a vertex), we define $\pi f'(P, t) = \tau(P, t)$ and $\lambda f'(P, t)$ to be a path in F'' joining $\lambda f'(P, 0)$ to $\lambda f'(P, 1)$. Suppose now $f'(x, t)$ is defined for x in any simplex preceding ζ , and is such that $\pi f'(x, t) = \tau(x, t)$. Then, if dimension $\zeta < h$, $f'(x, t)$ is defined on the boundary of $\zeta \times I$, and therefore, $\lambda f'(x, t)$ admits an extension over this prism. We extend $\pi f'(x, t)$ so as to be $\tau(x, t)$ over $\zeta \times I$. If, however, dimension $\zeta = h$, $f'(x, t)$ is defined only on $\partial \zeta \times I + \zeta \times 1$. As this set is a retract of $\zeta \times I$, $\lambda f'(x, t)$ admits an extension mapping $\zeta \times I$ in F'' . Then we set $\pi f'(x, t) = \tau(x, t)$. Now let $f'(x) = f'(x, 0)$ for $x \in M_2^h$. Then, for any ζ^{h+1} of M_2^h , $\lambda f'(\partial \zeta^{h+1})$ is homotopic in F'' to $\lambda f(\partial \tau \zeta^{h+1})$. Therefore $c_2(f', \zeta^{h+1}) = c_1(f, \tau \zeta^{h+1})$. Therefore, the cocycle $c_2^{h+1}(f')$ is the image of the cocycle $c_1^{h+1}(f)$ under the simplicial map τ . This completes the proof of (f).

Appendix I. Extensions of method

In some problems a tensor may already be defined on a submanifold L of M , and it may be required to extend it over M . Such a problem could arise in connection with a manifold M with regular boundary L . If M is subdivided so that L is a subcomplex, then as before f extends to $L + M^h$. The chain $c^{h+1}(f)$ lies in $M - L$ and is a cocycle in the open complex $M - L$. Then a necessary and sufficient condition that f on L can be extended to $L + M^{h+1}$ is that $c^{h+1}(f) \sim 0$ in $M - L$.

In those problems for which the characteristic cohomology class is zero and maps f of M^{h+1} can be defined, we are faced with the difficulty of extending such an f to M^{h+2} . We can, of course, proceed as before and define a chain $c^{h+2}(f)$ with coefficients in $\pi_{h+1}(F'')$. For this, it is necessary to know that F'' is $(h+1)$ -simple in the sense of Eilenberg,⁷ so that $\lambda f(\partial \sigma^{h+2})$ determines a unique element of $\pi_{h+1}(F'')$. It will follow as before that $c^{h+2}(f)$ is a cocycle; and $c^{h+2}(f) \sim 0$ is a sufficient condition for an f' defined on M^{h+2} to exist. However, it is not necessary; for the cohomology class of $c^{h+2}(f)$ may well vary from one f to another. Just how this cohomology class varies is not known. Some deeper analysis is required to resolve the difficulty.

Appendix II. Application to fibre spaces

In a paper by Hurewicz and the author [3], the notion of a fibre space was introduced as a generalization of the Whitney notion of fibre bundle. In those cases for which we know that F'' is compact, we can prove that M'' is a fibre space over M relative to π . This suggests that any fibre space X over a base space B relative to a mapping $\pi(X) = B$ determines in some sense a characteristic cohomology class in B . We propose to show that this is the case.

Assuming B to be arcwise connected, the fibres F may be deformed along

⁷ If $h > 1$, then $\pi_1(F'') = 0$, and F'' is i -simple for every i .

curves into others and these deformations induce isomorphisms between the $\pi_h(F)$ (where h is the smallest integer such that $\pi_h(F) \neq 0$). Therefore, the coefficient groups $\pi_h(F)$ are at hand as before.

A fibre space X need not be locally a product space over its base space B . Consequently, the $\pi f, \lambda f$ method must be replaced by a new mechanism. Such a one is provided by the following lemma.

LEMMA 3. *Let X be a fibre space over B relative to π . Let K be a complex, L a subcomplex, ψ a map of K in B , and ψ' a map of L in X such that $\pi\psi' = \psi$. Suppose ψ_t is a homotopy of $\psi(K)$ and ψ'_t is a homotopy of $\psi'(L)$ such that $\pi\psi'_t = \psi_t$, $0 \leq t \leq 1$. Then, if ψ'_1 admits an extension to K such that $\pi\psi'_1 = \psi_1$, this is also true of ψ' .*

The complex $U = L \times I + K \times 1$ is a deformation retract of $K \times I$. Let h_τ ($0 \leq \tau \leq 1$) be such a retraction. Since ψ'_1 gives a map of U in X such that $\pi\psi'_1 = \psi_1$ on U , the function $\psi'_1 h_1$ maps $K \times I$ in X such that $\pi\psi'_1 h_1 = \psi_1 h_1$. Since $\psi_1 h_1$ is homotopic to ψ_1 leaving U pointwise fixed, the covering homotopy [3] deforms $\psi'_1 h_1$ leaving U fixed into an extension of ψ'_1 to $K \times I$ such that $\pi\psi'_1 = \psi_1$. Then ψ'_0 is the required extension of ψ' .

Returning now to the problem of defining a characteristic cocycle, let K be a complex and ψ a map of K in B . We propose to show there is a map ψ' of K^h in X such that $\pi\psi' = \psi$. There is no difficulty in defining ψ' on K^0 . Suppose it has been properly defined on K^i and σ is an $(i+1)$ -simplex. Then ψ is defined on σ , and ψ' on $\partial\sigma$. Let h_t contract σ on itself to a point. Let $\psi_t = \psi h_t$, and let ψ'_t cover ψ_t on $\partial\sigma$. If $i < h$, the map ψ'_1 of $\partial\sigma$ in the fibre F over $\psi_1(\sigma)$ can be extended to a map of σ in F . Applying Lemma 3 gives an extension of ψ' over σ . Thus, ψ' can be defined over K^h .

If ψ' is a map of K^h in X such that $\pi\psi' = \psi$, we can define a chain $c^{h+1}(\psi')$ in K as follows. Let σ be an $(h+1)$ -simplex, and let h_t contract $\partial\sigma$ over σ to a point P_0 . Then the homotopy $\psi_t = \psi h_t$ of $\partial\sigma$ has a covering homotopy ψ'_t of ψ' . As ψ'_1 maps $\partial\sigma$ into the fibre F over $\psi(P_0)$, it defines an element of $\pi_h(F)$ which we denote by $c(\psi', \sigma^{h+1})$. Define $c^{h+1}(\psi') = \sum c(\psi', \sigma^{h+1}) \sigma^{h+1}$. Then we have:

THEOREM 5. *If X is a fibre space over B relative to π , K a complex, and ψ a map of K in B , then there exist maps ψ' of K^h in X such that $\pi\psi' = \psi$, where h is the smallest integer for which $\pi_h(F) \neq 0$. Any such map determines a cocycle $c^{h+1}(\psi')$ in K whose cohomology class c^{h+1} is independent of ψ' . It may be chosen arbitrarily in its class by an appropriate choice of ψ' . Therefore, in order that there be a map ψ' of K^{h+1} in X such that $\pi\psi' = \psi$, it is necessary and sufficient that the cohomology class $c^{h+1} = 0$. Furthermore, the class c^{h+1} is independent of the subdivision used in K .*

The proofs are like those of Theorem 4. They differ only in that covering homotopies and Lemma 3 are used in place of the $C \times F''$ construction.

A continuous finite cycle in B is composed of three things: a complex K , a map ψ of K in B , and a finite cycle Z in K . Let us consider $(h+1)$ -cycles with local coefficients in the character groups of the $\pi_h(F)$. Then $c^{h+1} \cdot Z^{h+1}$ is a real

number mod 1. If (K_1, ψ_1, Z_1) and (K_2, ψ_2, Z_2) are homologous continuous cycles, then by definition, there is a complex K_3 containing K_1, K_2 as subcomplexes, a map ψ_3 of K_3 in B agreeing with ψ_1, ψ_2 on K_1, K_2 , and a chain in K_3 whose boundary is $Z_1 - Z_2$. If a map ψ'_3 of K_3 in X is given such that $\pi\psi'_3 = \psi_3$, then $c_1^{h+1}(\psi'_3), c_2^{h+1}(\psi'_3)$ and $c_3^{h+1}(\psi'_3)$ are simultaneously defined, $c_1^{h+1}(\psi'_3)$ is the part of $c_3^{h+1}(\psi'_3)$ on K_1 , and $c_2^{h+1}(\psi'_3)$ the part on K_2 . Therefore, $c_1^{h+1} \cdot Z_1 = c_3^{h+1} \cdot Z_1$, $c_2^{h+1} \cdot Z_2 = c_3^{h+1} \cdot Z_2$. Since $Z_1 \sim Z_2$ in K_3 , $c_3^{h+1} \cdot Z_1 = c_3^{h+1} \cdot Z_2$. Thus, the real number mod 1, $c^{h+1} \cdot Z$, is independent of the representation of the homology class of Z . Thus, a map of the homology group $H^{h+1}(B)$ in the reals mod 1 is at hand. This map is homomorphic; for the sum of two cycles is represented by the abstract sum $K_1 + K_2$, a map $\psi = \psi_1$ on K_1 , $= \psi_2$ on K_2 , and the sum $Z_1 + Z_2$. The characteristic cocycle on $K_1 + K_2$ is the sum of the cocycles of K_1 and of K_2 . Therefore, $c^{h+1} \cdot (Z_1 + Z_2) = c_1^{h+1} \cdot Z_1 + c_2^{h+1} \cdot Z_2$. Summarizing, we have:

THEOREM 6. *Corresponding to the fibre space X over B , there is a characteristic cohomology class c^{h+1} in B , where h is the smallest integer such that $\pi_h(F) \neq 0$. c^{h+1} belongs to the cohomology group of characters of the homology group $H^{h+1}(B)$ based on continuous finite cycles with local coefficients in the character groups of the $\pi_h(F)$. If K is a complex, and ψ a map of K in B , then the image of c^{h+1} in K under ψ is the characteristic cohomology class of K relative to ψ (see Theorem 5).*

THE UNIVERSITY OF CHICAGO

BIBLIOGRAPHY

1. S. S. CAIRNS, *Triangulation of the manifold of class one*, Bull. Amer. Math. Soc., 41 (1935), pp. 549-552.
2. S. EILENBERG, *Cohomology and continuous mappings*, Annals of Math., 41 (1940), pp. 231-251.
3. W. HUREWICZ AND N. E. STEENROD, *Homotopy relations in fibre spaces*, Proc. Nat. Acad. Sci., 27 (1941), pp. 61-64.
4. W. HUREWICZ, *Beiträge zur Topologie der Deformationen I-IV*, Proc. Amsterdam Acad., 38 (1935), pp. 112 and 521, also 39 (1936), pp. 117 and 215.
5. E. STIEFEL, *Richtungsfelder und Fernparallelismus in n -dimensionalen Mannigfaltigkeiten*, Comm. Math. Helv., 8 (1936), pp. 3-51.
6. H. WHITNEY, *On the theory of sphere-bundles*, Proc. Nat. Acad. Sci., 26 (1940), pp. 148-153.
7. H. WHITNEY, *Differentiable manifolds*, Annals of Math., 37 (1936), pp. 645-680.

HOMOTOPY PROPERTIES OF THE REAL ORTHOGONAL GROUPS

BY GEORGE W. WHITEHEAD

(Received September 25, 1941)

1. Introduction

In this paper we propose to investigate the topological structure of the rotation group R_n of the n -sphere, with special emphasis on its homotopy properties. Of particular interest are the homotopy groups π_i of R_n . These groups, one for each dimension i , were first defined for a general space by Hurewicz [1].¹ Like the homology groups, they are topological invariants of a space; unlike the homology groups, however, no general method for computing them is known. Each space thus presents a problem in itself.

The computation of the groups $\pi_i(R_n)$ for $i \leq 5$ and all n will be carried out by the method of fibre mappings and covering homotopies developed by Hurewicz and Steenrod [2]. We shall make extensive use of the results of Freudenthal [3], Hopf [4], and Pontrjagin [5] on the homotopy groups of spheres.

The groups $\pi_i(R_n)$ are useful not only in the study of the homotopy properties of spheres, but also are used by Whitney in his theory of sphere-bundles [6, 7], where they appear as coefficient groups for certain co-cycle invariants.

Another application of our results appears in the theory of continuous vector fields over spheres. It is well known that no continuous field of unit vectors can be defined over the spheres of even dimension. Over the odd-dimensional spheres, however, one such vector field can always be defined; and if $n \equiv 3 \pmod{4}$ or $n \equiv 7 \pmod{8}$ it is possible to define three or seven independent vector fields, respectively, over the n -sphere S^n . These can be readily constructed by the use of the multiplication matrices for quaternions and Cayley numbers. For a general odd n , however, there is no known result on the maximum number of independent vector fields which can be defined over S^n .

In this paper the case $n \equiv 1 \pmod{4}$ is resolved as follows: *Any two vector fields over S^{4m+1} ($m = 0, 1, 2, \dots$) are somewhere dependent.* As a corollary to this result it is observed that *the tangent sphere-bundle of S^n is not simple if $n > 1$ and $n \equiv 1 \pmod{4}$.*

This investigation was carried out under the direction of Prof. N. E. Steenrod, to whom the author wishes to acknowledge his indebtedness for many valuable suggestions and criticisms.

2. Table of groups π_1 to π_5 of R_n

In this section the results of our computation of the homotopy groups $\pi_i(R_n)$ are exhibited, and a set of generators for these groups is given. Proofs will be deferred until Section 8.

¹ Numbers in square brackets refer to the bibliography at the end of the paper.

Let ∞ denote the free cyclic group, 2 the cyclic group of order two. If A and B are two abelian groups, $A + B$ denotes their direct sum. In terms of these notations, the groups $\pi_i(R_n)$ may be tabulated as follows:

	R_1	R_2	R_3	R_4	R_5	R_6	\dots	R_n	\dots
π_1	∞	2	2	2	2	2	\dots	2	\dots
π_2	0	0	0	0	0	0	\dots	0	\dots
π_3	0	∞	$\infty + \infty$	∞	∞	∞	\dots	∞	\dots
π_4	0	2	$2 + 2$	2	0	0	\dots	0	\dots
π_5	0	0	0	0	∞	0	\dots	0	\dots

The results of the first two rows were first obtained by Cartan [8].

A generator of $\pi_1(R_n)$ is given by the map of the circle $x_1^2 + x_2^2 = 1$ defined by

$$x \rightarrow \begin{pmatrix} x_1 & -x_2 & 0 \\ x_2 & x_1 & 0 \\ 0 & 0 & I_{n-1} \end{pmatrix},$$

where I_{n-1} is the $(n-1)$ -rowed identity matrix.

A generator of $\pi_3(R_n)$ ($n \geq 3$) is given by the map of the 3-sphere $\sum_{i=1}^4 x_i^2 = 1$

$$x \rightarrow \begin{pmatrix} x_1 & -x_2 & -x_3 & -x_4 & 0 \\ x_2 & x_1 & -x_4 & x_3 & 0 \\ x_3 & x_4 & x_1 & -x_2 & 0 \\ x_4 & -x_3 & x_2 & x_1 & 0 \\ 0 & 0 & 0 & 0 & I_{n-3} \end{pmatrix}.$$

The corner matrix is the linear transformation of Euclidean 4-space defined by multiplying every quaternion on the left by $x_1 + ix_2 + jx_3 + kx_4$.

The generator of $\pi_3(R_2)$ is the well-known double covering of R_2 by S^3 . Bordering this matrix with a 1 in the lower right hand corner and zeros elsewhere, we obtain the extra generator of $\pi_3(R_3)$.

To obtain the generator of $\pi_4(R_2)$, we map S^4 on S^3 essentially, and then map S^3 into R_2 by means of the generator of $\pi_3(R_2)$ given above. Such an essential map was constructed by Freudenthal [3]. In a similar manner we obtain generators for $\pi_4(R_3)$ and $\pi_4(R_4)$.

Finally, the generator of $\pi_5(R_5)$ is determined by the map of S^5 into R_5 given by

$$x \rightarrow \|\delta_{ij} - 2x_i x_j\| \cdot \begin{pmatrix} I_5 & 0 \\ 0 & -1 \end{pmatrix}$$

3. Preliminary notions

In this section we introduce notations and concepts which will occur throughout the paper. The relative and absolute homotopy groups of a space are introduced and two homomorphisms relating these groups are discussed.

Let points x in Euclidean $(n + 1)$ -space be referred to coördinates $(x_1, x_2, \dots, x_{n+1})$. The unit sphere $\sum x_i^2 = 1$ we denote by S^n . The *equatorial plane* $x_{n+1} = 0$ divides S^n into two hemispheres V_1^n and V_2^n defined by the inequalities $x_{n+1} \geq 0$ and $x_{n+1} \leq 0$ respectively. If $x = (x_1, x_2, \dots, x_{n+1}) \in S^n$, the *antipodal point* $(-x_1, -x_2, \dots, -x_{n+1})$ is denoted by \tilde{x} . We shall refer to the point $x^o = (0, 0, \dots, 1)$ as the *north pole*, and to its antipode \tilde{x}^o as the *south pole*.

The group R_n of all rotations of S^n may be represented as the group of all real square orthogonal matrices of order $n + 1$ with determinant $+1$. The subgroup of R_n consisting of all those rotations of S^n which leave the north pole fixed is isomorphic with the group R_{n-1} , and we shall denote the former group also by the symbol R_{n-1} .

Let Y be a topological space, F a closed subset of Y , and y_o a fixed point of F . Let \mathfrak{X} denote the space of all maps of V_1^n into Y which carry the boundary $\partial V_1^n = S^{n-1}$ into F and the north pole x^o of S^{n-1} into y_o . We introduce an equivalence relation in \mathfrak{X} as follows: two maps $f_1, f_2 \in \mathfrak{X}$ are said to be *equivalent* if they are homotopic, and during the homotopy S^{n-1} remains in F and x^o remains at y_o . In other words, two points of \mathfrak{X} are equivalent if they can be joined by an arc in \mathfrak{X} . The relation of equivalence is easily seen to be reflexive, symmetric, and transitive, and thus divides \mathfrak{X} into classes of equivalent maps, called *homotopy classes*. The homotopy class determined by a map f we denote by $\{f\}$; the set of all such homotopy classes by $\pi_n(Y, F)$. Hurewicz [1] has introduced an operation, called addition, in $\pi_n(Y, F)$, by means of which it becomes a group, the n^{th} *relative homotopy group of Y modulo F* . If the closed set F is specialized to consist only of the point y_o , the group $\pi_n(Y, y_o)$ so obtained is called the *absolute homotopy group* $\pi_n(Y)$. The latter group may also be defined by means of mappings of spheres into Y and we shall frequently find it convenient to use this definition.

We now introduce a homomorphism ω of $\pi_n(Y, F)$ into $\pi_{n-1}(F)$. This homomorphism is defined as follows: if $\alpha = \{f\} \in \pi_n(Y, F)$, then $\omega(\alpha)$ denotes the homotopy class of $\pi_{n-1}(F)$ determined by the map $f(S^{n-1}) \subset F$. Evidently $\{f\} = \{g\}$ implies $\omega(\{f\}) = \omega(\{g\})$. Thus ω maps $\pi_n(Y, F)$ into $\pi_{n-1}(F)$, and it follows from the definition of addition that ω is a homomorphism. Let $\pi_{no}(Y, F)$ denote the kernel of this homomorphism, $\pi_{n-1,o}(F)$ the image of $\pi_n(Y, F)$ under ω . Evidently $\pi_{no}(Y, F)$ consists of those homotopy classes determined by those relative n -cells in Y modulo F which are contractible into F , while $\pi_{n-1,o}(F)$ consists of the classes determined by those $(n - 1)$ -spheres in F which are homotopic to points in Y .

Since each element of $\pi_n(Y)$, considered as a set, is a subset of an element of $\pi_n(Y, F)$, we have a natural mapping ψ of $\pi_n(Y)$ into $\pi_n(Y, F)$, which is a homomorphism. It is not hard to show that $\psi(\{f\}) = 0$ if and only if some f' in the

class of f maps S^n into F . If $\{f_1\} = \{f_2\}$ and $\psi(\{f_1\}) = \psi(\{f_2\}) = 0$, then f'_1 is homotopic in Y to f'_2 . Thus f'_1 and f'_2 determine the same element of $\pi_n(F)/\pi_{no}(F)$. Conversely, if $f(S^n) \subset F$, then $\psi(\{f\}) = 0$. Hence the kernel of the homomorphism ψ is isomorphic to $\pi_n(F)/\pi_{no}(F)$. The image of $\pi_n(Y)$ under ψ is evidently the group $\pi_{no}(Y, F)$. We summarize these results in

THEOREM 1. *The natural homomorphic maps $\omega[\pi_n(Y, F)] \subset \pi_{n-1}(F)$ and $\psi[\pi_n(Y)] \subset \pi_n(Y, F)$ are related as follows: the kernel of the homomorphism ψ is isomorphic to $\pi_n(F)/\pi_{no}(F)$, while the image of $\pi_n(Y)$ under ψ is the group $\pi_{n,o}(Y, F)$.*

4. Homotopy relations in compact Lie groups

Let G be a topological group, H a closed subgroup of G , and $B = G/H$ the space of left (or right) cosets of H in G . Then there is a natural mapping π of G onto B defined as follows: for every $g \in G$, $\pi(g)$ is the coset of B containing g .

If G is a compact Lie group, then π is a fibre map in the sense of Hurewicz and Steenrod [2]. A slicing function can be defined as follows: a plane of maximum dimension independent of the tangent plane to H at the identity 1 meets each coset b in a sufficiently small neighborhood U of $b_o = \pi(1)$ just once. We denote this point by $\phi(1, b)$. Then if $g^{-1}b \in U$, let $\phi(g, b) = g\phi(1, g^{-1}b)$. Evidently ϕ has all the required properties of a slicing function.

The following theorem will be useful in our discussion of the rotation groups:

THEOREM 2. *If G is a topological group and B is the space of left (or right) cosets of a closed subgroup H of G , and if there exists a map $f(B) \subset G$ such that $\pi f(b) = b$, then G is homeomorphic with the product space $H \times B$.*

We may suppose B is a space of left cosets. Let $f'(b) = f(b) \cdot [f(b_o)]^{-1}$; then $\pi f'(b) = \pi f(b) = b$, and $f'(b_o) = 1$. We then set up the homeomorphism by means of two maps $p(G) = H \times B$ and $q(H \times B) = G$ defined as follows:

$$\begin{aligned} p(g) &= [g^{-1} \cdot f'(\pi g), \pi g] & (g \in G), \\ q(h, b) &= f'(b) \cdot h^{-1} & (h \in H, b \in B). \end{aligned}$$

Then

$$\begin{aligned} p[q(h, b)] &= \{h[f'(b)]^{-1}f'(b), b\} = (h, b), \\ q[p(g)] &= f'(\pi g) \cdot [f'(\pi g)]^{-1} \cdot g = g, \end{aligned}$$

and both maps are continuous. Hence G and $H \times B$ are homeomorphic.

5. Slicing functions for $R_n \rightarrow R_n/R_{n-1}$

In this section the results of Section 4 are applied to the special case $G = R_n$, $H = R_{n-1}$, and an explicit slicing function is constructed. The special cases $n = 1, 3, 7$ are treated separately, and for these values of n Theorem 2 is applied.

Let us consider the mapping $\pi(R_n) = S^n$ defined by $\pi(r) = r(x^o)$ ($r \in R_n$). As shown by Hurewicz and Steenrod [2], π is a fibre map of R_n into S^n , the fibres being the left cosets of R_{n-1} in R_n . In terms of the matrix representation of R_n , $\pi(r)$ is the last column of the matrix r .

We now define the slicing function promised above: if $x \neq \tilde{x}^0$, let $\phi(I, x)$ be the rotation carrying x^0 along a great circle into x , and leaving the $(n-2)$ -sphere orthogonal to this great circle fixed. In terms of matrices

$$(1) \quad \phi(I, x) = A_n(x) = \left\| \delta_{ij} - \frac{(x_i + \delta_{i,n+1})(x_j + \delta_{j,n+1})}{x_{n+1} + 1} \right\| \cdot \left\| \begin{matrix} I_n & 0 \\ 0 & -1 \end{matrix} \right\|$$

$$(i, j = 1, \dots, n+1),$$

where I_n denotes the n -rowed identity matrix. Then $\phi(r, x)$ is defined as in Section 2. Evidently it is impossible to extend $\phi(I, x)$ so as to be defined and continuous over all of S^n . We shall show later that there is no slicing function with this property for a general n .

For the dimensions 1, 3, and 7, however, it is possible to define such a slicing function. Let \mathfrak{A}_{n+1} ($n = 1, 3, 7$) denote the algebras of complex numbers, quaternions, and Cayley numbers, respectively, over the field of real numbers. By means of these algebras a multiplication $x \cdot y$ of points of Euclidean $(n+1)$ -space is defined. This multiplication has the property that $\|x \cdot y\| = \|x\| \cdot \|y\|$, where $\|x\|^2 = \sum x_i^2$ is the square of the distance of the point x from the origin. Hence S^n is closed under multiplication and for $x, y \in S^n$ we have $x \cdot y = B_n(x) \cdot y$, where $B_n(x) \in R_n$, and, if the coordinate system is chosen so that x^0 is the unit of the algebra, $B_n(x^0) = I$, $\pi B_n(x) = x$. Since $B_n(x)$ is defined for all $x \in S^n$, we have

THEOREM 3. For $n = 1, 3, 7$, R_n is a product space $R_{n-1} \times S^n$.

Since the i^{th} homotopy group of a product space $X \times Y$ is the direct sum of the i^{th} homotopy groups of X and Y , we have

COROLLARY. For $n = 1, 3, 7$, $\pi_i(R_n)$ is the direct sum $\pi_i(R_{n-1}) + \pi_i(S^n)$; in particular, $\pi_{n-1}(R_{n-1}) = \pi_{n-1}(R_n)$.

6. The canonical map of S^n in R_n

We now introduce a mapping C_n of S^n into R_n which plays an important role in the following discussion. We shall refer to it as the *canonical map*. It is proved that this map is contractible into R_{n-1} if n is even; while for n odd it is not so contractible. The canonical map is then used to construct a generator for $\pi_n(R_n, R_{n-1})$.

In order to define the map C_n , let $\theta(x)$ ($x \in S^n$) denote the angular distance from x^0 to x , and let x' be the point in the great circle through x^0 and x with $\theta(x') = 2\theta(x)$. Let $C_n(x)$ ($x \neq \tilde{x}^0$) be the rotation which carries x^0 along a great circle into x' and leaves the orthogonal $(n-2)$ -sphere fixed; and let $C_n(\tilde{x}^0) = I$. Evidently

$$(2) \quad C_n(x) = [A_n(x)]^2 = \left\| \delta_{ij} - 2x_i x_j \right\| \cdot \left\| \begin{matrix} I_n & 0 \\ 0 & -1 \end{matrix} \right\| \quad (i, j = 1, \dots, n+1).$$

We observe that C_n , unlike A_n , is defined and continuous over all of S^n . Furthermore, antipodal points have the same image, while distinct pairs of antipodal points have distinct images. Thus the image of S^n in R_n is a projective n -space.

Let $g_n(S^n) = S^n$ denote the projection πC_n of the canonical map. If $g_{n,i}(x)$ is the i^{th} coordinate of $g(x)$, we have $g_{n,i}(x) = 2x_i x_{n+1} - \delta_{i,n+1}$ ($i = 1, \dots, n+1$).

THEOREM 4. *If n is even, g_n has degree zero; if n is odd, g_n has degree two.*

For g_n maps the equator S^{n-1} into \tilde{x}^0 and maps $V_1^n - S^{n-1}$ topologically on $S^n - \tilde{x}^0$; in fact, g_n can be obtained by a homotopy of S^n on itself in which each point moves along the great circle joining it to the north pole. Thus g_n maps V_1^n on S^n with degree 1. Furthermore, g_n maps V_2^n on S^n with degree $(-1)^{n+1}$; for $g_n(x) = g_n(\tilde{x})$ ($x \in V_2^n$) and the antipodal transformation $x \rightarrow \tilde{x}$ has degree $(-1)^{n+1}$. Hence g_n maps S^n on itself with degree $1 + (-1)^{n+1}$.

If n is even, g_n has degree zero, and hence is homotopic to a point. We shall give a homotopy of g_n which will be useful in a later section. This homotopy $g_n(x, t)$ is given by the equations

$$\begin{aligned} g_{n,2i-1}(x, t) &= 2\{(1-t)x_{2i-1}x_{n+1} + [t(1-t)]^{\frac{1}{2}}x_{2i}\}, \\ (3) \quad g_{n,2i}(x, t) &= 2\{(1-t)x_{2i}x_{n+1} - [t(1-t)]^{\frac{1}{2}}x_{2i-1}\} \quad (i = 1, \dots, n/2), \\ g_{n,n+1}(x, t) &= 1 - 2(1-t)(1-x_{n+1}^2). \end{aligned}$$

It is easy to verify that $g_n(x, t)$ contracts $g_n(S^n)$ over S^n into x^0 .

If n is odd, g_n has degree two. We shall give a deformation of g_n into a second map g'_n of degree two. The latter map is defined by the equations

$$\begin{aligned} g'_{n,i}(x) &= x_i \quad (i = 1, \dots, n-1); \\ g'_{n,n}(x) &= \frac{2x_n x_{n+1}}{(x_n^2 + x_{n+1}^2)^{\frac{1}{2}}}, \\ (4) \quad g'_{n,n+1}(x) &= \frac{x_{n+1}^2 - x_n^2}{(x_n^2 + x_{n+1}^2)^{\frac{1}{2}}} \quad (x_n^2 + x_{n+1}^2 \neq 0), \\ g'_{n,n}(x) &= g'_{n,n+1}(x) = 0 \quad (x_n = x_{n+1} = 0). \end{aligned}$$

It is easily seen that g'_n is defined and continuous over all of S^n . The homotopy $g_n(x, t)$ of g_n over S^n into g'_n is given by

$$\begin{aligned} g_{n,2i-1}(x, t) &= tx_{2i-1} + [t(1-t)]^{\frac{1}{2}}x_{2i} + 2x_{n+1} \frac{(1-t)x_{2i-1} - [t(1-t)]^{\frac{1}{2}}x_{2i}}{\{1-t[1-(x_n^2 + x_{n+1}^2)]\}^{\frac{1}{2}}}, \\ g_{n,2i}(x, t) &= tx_{2i} - [t(1-t)]^{\frac{1}{2}}x_{2i-1} + 2x_{n+1} \frac{(1-t)x_{2i} + [t(1-t)]^{\frac{1}{2}}x_{2i-1}}{\{1-t[1-(x_n^2 + x_{n+1}^2)]\}^{\frac{1}{2}}}, \\ (5) \quad & \left(i = 1, \dots, \frac{n-1}{2}\right), \end{aligned}$$

$$\begin{aligned} g_{n,n}(x, t) &= \frac{2x_n x_{n+1}}{\{1-t[1-(x_n^2 + x_{n+1}^2)]\}^{\frac{1}{2}}}, \\ g_{n,n+1}(x, t) &= \frac{2x_{n+1}^2}{\{1-t[1-(x_n^2 + x_{n+1}^2)]\}^{\frac{1}{2}}} - \{1-t[1-(x_n^2 + x_{n+1}^2)]\}^{\frac{1}{2}}. \end{aligned}$$

Although $g_n(x, t)$ is not defined everywhere for $t = 1$, it is easily verified that $\lim_{t \rightarrow 1} g_n(x, t) = g'_n(x)$ uniformly in x .

We now use the canonical map to construct a generator of $\pi_n(R_n, R_{n-1})$. We shall take as fixed reference points for this group the north pole of S^{n-1} and the identity matrix $I \in R_{n-1}$. Since ([2], Theorem 2) $\pi_n(R_n, R_{n-1})$ is isomorphic to $\pi_n(S^n)$ under the map $\pi(R_n) = S^n$, a generator of the former group is represented by any relative n -cell in R_n modulo R_{n-1} which projects into S^n with degree ± 1 . To define such a relative n -cell, we observe that C_n maps V_1^n into R_n and S^{n-1} into the coset \tilde{R}_{n-1} opposite to R_{n-1} and projects into S^n with degree 1. Hence the map $D_n(x) = C_n(x) \cdot \begin{vmatrix} I_{n-1} & 0 \\ 0 & -I_2 \end{vmatrix} (x \in V_1^n)$ defines a relative cell in R_n modulo R_{n-1} , and it is easily verified that $D_n(x^0) = I$, while $d_n(x) = \pi D_n(x) = \tilde{g}_n(x)$ has degree $(-1)^{n+1}$. Hence D_n represents the required generator.

Since $\pi_{n-1}(R_n, R_{n-1}) = \pi_{n-1}(S^n) = 0$, it follows from the results of Section 3 that $\pi_{n-1}(R_n) = \pi_{n-1}(R_{n-1})/\pi_{n-1,o}(R_{n-1})$. Thus $\pi_{n-1}(R_n)$ is a factor group of $\pi_{n-1}(R_{n-1})$, the kernel of the homomorphism being the group $\pi_{n-1,o}(R_{n-1})$. But $\pi_{n-1,o}(R_{n-1}) = \omega[\pi_n(R_n, R_{n-1})]$; hence a generator of $\pi_{n-1,o}(R_{n-1})$ is given by the map ωD_n , which is easily shown to be the canonical map C_{n-1} . Hence

THEOREM 5. *The kernel of the homomorphism $\pi_{n-1}(R_{n-1}) \rightarrow \pi_{n-1}(R_n)$ is the subgroup of the former group generated by the canonical map.*

7. On the possibility of sectioning the cosets of R_{n-1} in R_n

In this section the following question is considered: Is there an n -sphere in R_n which projects into S^n with degree 1? It is shown that this question can be answered in the negative for certain values of n . For other dimensions the question remains open.

The first step toward the solution of this problem appears in

THEOREM 6. *The following conditions are equivalent:*

- 1) *there is a map $F(S^n) \subset R_n$ such that πF has degree one;*
- 2) *R_n can be represented as a product space $R_{n-1} \times S^n$;*
- 3) *the homomorphism $\pi_{n-1}(R_{n-1}) \rightarrow \pi_{n-1}(R_n)$ is an isomorphism;*
- 4) *the canonical map of S^{n-1} into R_{n-1} is homotopic to a point in R_{n-1} .*

The first condition implies the second. For let $F(S^n) \subset R_n$ be such that πF has degree one. Since πF is homotopic to the identity, it follows from the covering homotopy theorem ([2], Theorem 1) that F is homotopic to a map $F'(S^n) \subset R_n$ such that $\pi F'(x) = x$. Then by Theorem 2, $R_n = R_{n-1} \times S^n$.

That the second condition implies the third we have observed in the proof of the Corollary to Theorem 3; that the third implies the fourth follows from Theorem 5.

The fourth implies the first; for during the homotopy of C_{n-1} to a point a relative n -cell is swept out in R_{n-1} whose boundary is the $(n-1)$ -sphere defined by the canonical map. But the map D_n defines a relative n -cell in R_n with the same boundary as the first one. Joining these two cells by identifying corre-

sponding points on the boundaries, we obtain a sphere in R_n whose projection has the same degree $(-1)^{n+1}$ as that of d_n . The required map is then easily constructed.

Theorem 6 enables us to answer immediately the question posed above for the case when n is even. For suppose that n is even and suppose that such a map exists. Then by the fourth condition in Theorem 6, the canonical map C_{n-1} is homotopic to a point in R_{n-1} . Hence $g_{n-1} = \pi C_{n-1}$ is inessential. But we have already shown that g_{n-1} has degree two. This contradiction completes the proof of

THEOREM 7. *If n is even, there is no map $F(S^n) \subset R_n$ such that πF has degree one.*

We now turn to the much more difficult case where n is odd. A partial answer to the question is obtained in

THEOREM 8. *If $n > 1$ and $n \equiv 1 \pmod{4}$, there is no map $F(S^n) \subset R_n$ such that πF has degree one.*

The proof may be outlined as follows:

1) Since for n odd, g_{n-1} is homotopic to a point, it follows that C_{n-1} is homotopic to a map $G_{n-1}(S^{n-1}) \subset R_{n-2}$. Such a homotopy is exhibited, and it is proved that $k_{n-1} = \pi G_{n-1}$ is essential if $n \equiv 1 \pmod{4}$, and inessential if $n \equiv 3 \pmod{4}$.

2) A generator $P_n(V_1^n) \subset R_{n-1}$ of the group $\pi_n(R_{n-1}, R_{n-2})$ is constructed for all $n \geq 5$, and the projection p_{n-1}^* of the map $\omega P_n(S^{n-1}) \subset R_{n-2}$ is shown to be inessential.

When these steps have been established, the proof of the theorem may be completed as follows: Suppose that $n > 1$ and $n \equiv 1 \pmod{4}$ and suppose that the theorem is not true. Then C_{n-1} , and consequently G_{n-1} , is homotopic to a point in R_{n-1} . Since G_{n-1} is not homotopic to a point in R_{n-2} , the deformation of G_{n-1} defines a relative n -cell in R_{n-1} modulo R_{n-2} , and hence an essential element of $\pi_n(R_{n-1}, R_{n-2})$. This group has been shown by Freudenthal [3] to be the cyclic group of period 2 if $n \geq 4$. Hence ωP_n and G_{n-1} are homotopic in R_{n-2} . But p_{n-1}^* and k_{n-1} are not homotopic, a contradiction.

LEMMA. *If n is odd, the canonical map C_{n-1} is homotopic in R_{n-1} to a map $G_{n-1}(S^{n-1}) \subset R_{n-2}$, whose projection into S^{n-2} is essential if $n \equiv 1 \pmod{4}$ and inessential if $n \equiv 3 \pmod{4}$.*

Let E^n denote the closed n -cell bounded by the unit sphere S^{n-1} in Euclidean n -space. H^n denotes the upper half $x_n \geq 0$ of E^n . Let points $y \in H^n$ be represented by coördinates (x, r) where $x \in V_1^{n-1}$ is the central projection of the point y on V_1^{n-1} , and r is the distance of the point y from the origin. Let

$$H(r) = \begin{vmatrix} c & s & 0 & 0 & \cdots & 0 \\ -s & c & 0 & 0 & \cdots & 0 \\ 0 & 0 & c & s & \cdots & 0 \\ 0 & 0 & -s & c & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \end{vmatrix}$$

where $c = 1 - 2r^2$, $s = 2r(1 - r^2)^{\frac{1}{2}}$ ($0 \leq r \leq 1$). Let

$$(6) \quad G(y) = G(x, r) = C_{n-1}(x) \cdot H(r) \cdot C_{n-1}(x)^{-1} \cdot \begin{vmatrix} -I_{n-1} & 0 \\ 0 & 1 \end{vmatrix} \quad (x \in V_1^{n-1}; 0 \leq r \leq 1)s$$

map H^n into R_{n-1} . Evidently $G(x) = C_{n-1}[g_{n-1}(x)]$ for $x \in V_1^{n-1}$, while G map, the equatorial plane into R_{n-2} and S^{n-2} into a single point. Deform V_1^{n-1} over H^n into E^{n-1} by letting each point move along the perpendicular joining it to the plane $x_n = 0$. The image of G under this homotopy gives a homotopy of $C_{n-1}[g_{n-1}(x)]$ to a map $K(V_1^{n-1}) \subset R_{n-2}$. Since under the homotopy S^{n-2} remains fixed, K maps S^{n-2} into a single point. Evidently $K(x)$ can be represented in the form $K(x) = G_{n-1}[g_{n-1}(x)]$, where $G_{n-1}(S^{n-1}) \subset R_{n-2}$ is defined and continuous over all of S^{n-1} , and G_{n-1} is homotopic in R_{n-1} to C_{n-1} .

By multiplying out the matrices in (6) and computing the homotopy, we find that the map $k_{n-1} = \pi G_{n-1}$ of S^{n-1} into S^{n-2} is represented by the equations

$$(7) \quad \begin{aligned} y_{2i-1} &= \frac{2(x_{2i-1}x_{n-2} + x_{2i}x_{n-1})}{(1 - x_n^2)^{\frac{1}{2}}}, \\ y_{2i} &= \frac{2(x_{2i}x_{n-2} - x_{2i-1}x_{n-1})}{(1 - x_n^2)^{\frac{1}{2}}} \quad \left(i = 1, \dots, \frac{n-3}{2}\right), \\ y_{n-2} &= \frac{2(x_{n-2}^2 + x_{n-1}^2)}{(1 - x_n^2)^{\frac{1}{2}}} - (1 - x_n^2)^{\frac{1}{2}}, \\ y_{n-1} &= x_n \quad (1 - x_n^2 \neq 0); \\ y_i &= 0 \quad (i = 1, \dots, n-2), \quad y_{n-1} = x_n \quad (1 - x_n^2 = 0). \end{aligned}$$

It follows easily that k_{n-1} is defined and continuous over all of S^{n-1} and maps S^{n-1} on S^{n-2} . In order to investigate this map, let us consider the map $m_{n-1}(S^{n-2}) = S^{n-3}$ obtained by setting $x_n = y_{n-1} = 0$ in (7). This map can be studied more easily in complex coördinates. Let $z_j = x_{2j-1} + ix_{2j}$, $w_j = y_{2j-1} + iy_{2j}$ ($j = 1, \dots, (n-1)/2 = k$), where $w_k = y_{n-2}$ is real. Then S^{n-2} and S^{n-3} are given by the equations $\sum z_j \bar{z}_j = 1$, $\sum w_j \bar{w}_j = 1$, respectively. In terms of these coördinates the equations representing the map m_{n-1} take the form

$$(8) \quad \begin{aligned} w_j &= 2z_j \bar{z}_k \quad (j = 1, \dots, k-1), \\ w_k &= 2z_k \bar{z}_k - 1. \end{aligned}$$

If the complex coördinates occurring in (8) are formally replaced by real ones, the map g_{k-1} is obtained. In equations (3) and (5) we have constructed homotopies of g_n for n even and odd, respectively. The functions defining these homotopies can be extended so as to be defined for complex coördinates, as follows: if k is odd, i.e., if $n \equiv 3 \pmod{4}$, let

$$w_{2j-1} = 2\{(1-t)z_{2j-1}\bar{z}_k + [t(1-t)]^{\frac{1}{2}}\bar{z}_{2j}\},$$

$$(9) \quad \begin{aligned} w_{2j} &= 2\{(1-t)z_{2j}\bar{z}_k - [t(1-t)]^{\frac{1}{2}}\bar{z}_{2j-1}\} & (j = 1, \dots, \frac{k-1}{2}), \\ w_k &= 1 - 2(1-t)(1 - z_k\bar{z}_k). \end{aligned}$$

Evidently the homotopy (9) deforms m_{n-1} over S^{n-3} to a point. Hence m_{n-1} , and consequently also k_{n-1} , is homotopic to a point. On the other hand, if k is even, i.e., if $n \equiv 1 \pmod{4}$, let us consider the map $m'_{n-1}(S^{n-2}) \subset S^{n-3}$ defined by the equations

$$(10) \quad \begin{aligned} w_j &= z_j & (j = 1, \dots, k-2), \\ w_{k-1} &= \frac{2z_{k-1}\bar{z}_k}{(z_{k-1}\bar{z}_{k-1} + z_k\bar{z}_k)^{\frac{1}{2}}}, \\ w_k &= \frac{2z_k\bar{z}_k}{(z_{k-1}\bar{z}_{k-1} + z_k\bar{z}_k)^{\frac{1}{2}}} - (z_{k-1}\bar{z}_{k-1} + z_k\bar{z}_k)^{\frac{1}{2}} & (z_{k-1}\bar{z}_{k-1} + z_k\bar{z}_k \neq 0); \\ w_{k-1} &= w_k = 0 & \text{for } z_{k-1} = z_k = 0. \end{aligned}$$

Setting $w_j = z_j = 0$ ($j = 1, \dots, k-2$) in (10) and writing the result in real coördinates, we obtain a map of S^3 on S^2 of Hopf invariant ± 1 (cf. Hopf [4], §5). It follows from the results of Freudenthal [3] that m'_{n-1} is essential.

We shall now define a deformation of m_{n-1} to m'_{n-1} (for $n \equiv 1 \pmod{4}$) by extending the functions in (5) so as to be defined for complex coördinates, as follows: let

$$(11) \quad \begin{aligned} w_{2j-1} &= tz_{2j-1} + [t(1-t)]^{\frac{1}{2}}\bar{z}_{2j} + 2 \frac{(1-t)z_{2j-1}\bar{z}_k - [t(1-t)]^{\frac{1}{2}}z_k\bar{z}_{2j}}{\{1 - t[1 - (z_{k-1}\bar{z}_{k-1} + z_k\bar{z}_k)]\}^{\frac{1}{2}}}, \\ w_{2j} &= tz_{2j} - [t(1-t)]^{\frac{1}{2}}\bar{z}_{2j-1} + 2 \frac{(1-t)z_{2j}\bar{z}_k + [t(1-t)]^{\frac{1}{2}}z_k\bar{z}_{2j-1}}{\{1 - t[1 - (z_{k-1}\bar{z}_{k-1} + z_k\bar{z}_k)]\}^{\frac{1}{2}}}, \\ & & (j = 1, \dots, \frac{k-2}{2}), \\ w_{k-1} &= \frac{2z_{k-1}\bar{z}_k}{\{1 - t[1 - (z_{k-1}\bar{z}_{k-1} + z_k\bar{z}_k)]\}^{\frac{1}{2}}}, \\ w_k &= \frac{2z_k\bar{z}_k}{\{1 - t[1 - (z_{k-1}\bar{z}_{k-1} + z_k\bar{z}_k)]\}^{\frac{1}{2}}} - \{1 - t[1 - (z_{k-1}\bar{z}_{k-1} + z_k\bar{z}_k)]\}^{\frac{1}{2}}. \end{aligned}$$

Although this map is not defined everywhere for $t = 1$, it is readily verified that as $t \rightarrow 1$, the functions in (11) converge to the corresponding ones in (10) uniformly in z . Thus m_{n-1} and m'_{n-1} are homotopic, so that m_{n-1} , and consequently also k_{n-1} , is essential. This completes the proof of the Lemma.

As indicated above, the next step in the proof is to construct a generator of the group $\pi_n(R_{n-1}, R_{n-2})$. This group is isomorphic with $\pi_n(S^{n-1})$ under the projection π ([2], Theorem 2). If $n \geq 4$ we may take as generator of the latter group the map $p_n(S^n) = S^{n-1}$ defined by

$$p_{n,i}(x) = x_i \quad (i = 1, \dots, n-4),$$

$$\begin{aligned}
 p_{n,n-3}(x) &= \frac{2(x_{n-3}x_{n-1} + x_{n-2}x_n)}{(x_n^2 + x_{n-1}^2 + x_{n-2}^2 + x_{n-3}^2)^{\frac{1}{2}}}, \\
 p_{n,n-2}(x) &= \frac{2(x_{n-2}x_{n-1} - x_{n-3}x_n)}{(x_n^2 + x_{n-1}^2 + x_{n-2}^2 + x_{n-3}^2)^{\frac{1}{2}}}, \\
 p_{n,n-1}(x) &= \frac{x_n^2 + x_{n-1}^2 - x_{n-2}^2 - x_{n-3}^2}{(x_n^2 + x_{n-1}^2 + x_{n-2}^2 + x_{n-3}^2)^{\frac{1}{2}}}, \\
 p_{n,n}(x) &= x_{n+1} \quad (x_n^2 + x_{n-1}^2 + x_{n-2}^2 + x_{n-3}^2 \neq 0), \\
 p_{n,n-3}(x) &= p_{n,n-2}(x) = p_{n,n-1}(x) = 0 \quad (x_n = x_{n-1} = x_{n-2} = x_{n-3} = 0).
 \end{aligned}
 \tag{12}$$

Evidently p_n maps V_1^n into V_1^{n-1} and S^{n-1} into S^{n-2} . The latter map is essential; in fact, it represents a generator of $\pi_{n-1}(S^{n-2})$. Let $P_n(x) = D_{n-1}[p_n(x)]$ ($x \in V_1^n$). Then P_n represents the desired generator of $\pi_n(R_{n-1}, R_{n-2})$; for $\pi P_n(x) = d_{n-1}[p_n(x)]$, and the latter map represents a generator of $\pi_n(S^{n-1}, x^0) = \pi_n(S^{n-1})$.

Let $P_n^* = \omega P_n$, and $p_{n-1}^* = \pi P_n^*$ its projection into S^{n-2} . Since p_n maps S^{n-1} into S^{n-2} , and since $\omega D_{n-1} = C_{n-2}$, we have $P_n^*(x) = C_{n-2}[p_n(x)]$, and hence $p_{n-1}^*(x) = g_{n-2}[p_n(x)]$ ($x \in S^{n-1}$). Let α denote the element of $\pi_{n-2}(S^{n-2})$ determined by the identity map, β the element of $\pi_{n-1}(S^{n-2})$ defined by the map $p_n(S^{n-1}) = S^{n-2}$. Since $\{g_{n-2}\} = 0$ or 2α according as n is even or odd, we have $\{p_{n-1}^*\} = 0$ or $4\beta^2$ respectively. But Freudenthal [3] has proved that $2\beta = 0$ if $n \geq 5$; hence for all $n \geq 5$ we have $\{p_{n-1}^*\} = 0$, so that p_{n-1}^* is inessential. This completes the proof of the theorem.

8. Computation of the homotopy groups

We are now in a position to establish the results exhibited in Section 2. The generators which we shall compute here will differ slightly from those already exhibited; however, they may be easily seen to be homotopic.

We have already observed that $\pi_n(R_{n+1})$ is a homomorphic image of $\pi_n(R_n)$. In a similar manner we can prove (cf. [2], Theorem 5) that $\pi_n(R_{n+k})$ ($k = 1, 2, \dots$) is isomorphic with $\pi_n(R_{n+1})$. Another useful result is the following:

THEOREM 9. *If n is even, $\pi_n(R_n)$ is a factor group of $\pi_n(R_{n-1})$; the kernel of the homomorphism is the group $\pi_{n,0}(R_{n-1})$.*

Since $\{D_n\}$ is a generator of $\pi_n(R_n, R_{n-1})$ and since $\omega\{D_n\} = \{C_{n-1}\} \neq 0$ for n even, it follows that $\pi_{n,0}(R_n, R_{n-1}) = 0$. Hence $\pi_n(R_n) = \pi_n(R_{n-1})/\pi_{n,0}(R_{n-1})$.

In order to compute $\pi_1(R_n)$, we first observe that, since R_1 is homeomorphic to S^1 , its fundamental group is the free cyclic group generated by any map of S^1 on R_1 of degree 1. Such a generator α_1 is given by the map

$$B_1(x) = \left\| \begin{array}{cc} x_2 & x_1 \\ -x_1 & x_2 \end{array} \right\|.$$

Since R_2 is homeomorphic with projective 3-space P^3 , it follows that $\pi_1(R_2)$

* This follows easily from Theorem II b' of [4].

is the cyclic group of order two. But $\pi_1(R_2)$ is a factor group of $\pi_1(R_1)$; hence for a generator of $\pi_1(R_2)$ we may take the generator α_1 for $\pi_1(R_1)$ subject to the condition $2\alpha_1 = 0$ in R_2 . The same result holds for $\pi_1(R_n)$.

Since all the higher homotopy groups of S^1 vanish, the same holds for R_1 . In particular, $\pi_2(R_1) = 0$. Then, by Theorem 9, $\pi_2(R_n) = 0$ for all n .

The higher homotopy groups of R_2 are isomorphic to those of its covering space S^3 . In particular, $\pi_3(R_2)$ is the free cyclic group, and a generator α_3 is represented by the covering map $H(S^3) = R_2$ given by the equation

$$H(x) = \begin{vmatrix} x_4^2 - x_3^2 - x_2^2 + x_1^2 & 2(x_1x_2 - x_3x_4) & 2(x_1x_3 + x_2x_4) \\ 2(x_1x_3 + x_2x_4) & x_4^2 - x_3^2 + x_2^2 - x_1^2 & 2(x_2x_3 - x_1x_4) \\ 2(x_1x_3 - x_2x_4) & 2(x_1x_4 + x_2x_3) & x_4^2 + x_3^2 - x_2^2 - x_1^2 \end{vmatrix}$$

We observe that $\pi H(x)$ maps S^3 on S^2 with Hopf invariant ± 1 .

Since R_3 is the product space of R_2 and the quaternion subgroup Q^3 , the homotopy groups of R_3 are the direct sums of the groups of the same dimension of R_2 and $Q^3 = S^3$. In particular, $\pi_3(R_3) = \pi_3(R_2) + \pi_3(S^3)$ is a free group with two generators. For one of these generators we may take the generator α_3 of $\pi_3(R_2)$; the second generator β_3 is determined by the quaternion matrix

$$B_3(x) = \begin{vmatrix} x_4 & x_3 & -x_2 & x_1 \\ -x_3 & x_4 & x_1 & x_2 \\ -x_2 & -x_1 & x_4 & x_3 \\ -x_1 & -x_2 & -x_3 & x_4 \end{vmatrix} \quad (x \in S^3).$$

To compute $\pi_3(R_4) = \pi_3(R_3)/\pi_{3,o}(R_3)$, we observe that

$$C_3(x) = B_3[g_3(x)] \cdot \begin{vmatrix} H(x) & 0 \\ 0 & 1 \end{vmatrix}$$

so that $\{C_3\} = 2\beta_3 + \alpha_3 = 0$ in R_4 . Hence $\pi_3(R_4)$ is a free cyclic group with the one generator β_3 , and the same is true of $\pi_3(R_n)$ ($n \geq 4$).

We now consider the groups π_4 . Freudenthal [3] and Pontrjagin [5] have proved that $\pi_4(S^3)$ is the cyclic group of order two, a generator being determined by the map $p_4(S^4) = S^3$. Hence a generator α_4 of $\pi_4(R_2)$ is determined by the map $H p_4$. The group $\pi_4(R_3)$ is the direct sum of two cyclic groups of order two; for generators we may take α_4 and $\beta_4 = \{B_3 p_4\}$.

We have observed in the proof of Theorem 8 that $\pi_{4,o}(R_3)$ is generated by $\omega\{P_3\} = \{P_3^*\}$. But

$$P_3^*(x) = B_3[p_4^*(x)] \cdot \begin{vmatrix} H[p_4(x)] & 0 \\ 0 & 1 \end{vmatrix}$$

and $\{p_4^*\} = 0$, so that $\{P_3^*\} = \alpha_4 = 0$ in R_4 . Hence $\pi_4(R_4)$ is the cyclic group of order two generated by β_4 .

We have proved in Theorem 8 that $\{C_4\} \neq 0$ in R_4 . Hence $\pi_{4,o}(R_4) = \pi_4(R_4)$, so that $\pi_4(R_5) = \pi_4(R_n) = 0$ ($n \geq 5$).

We conclude by computing the five-dimensional homotopy groups of R_n . Pontrjagin [5] has proved that $\pi_5(S^3) = 0$; hence $\pi_5(R_2) = \pi_5(R_3) = 0$. Hence $\pi_5(R_4) = \pi_{5,o}(R_4, R_3) = 0$ since $\pi_{4,o}(R_3) \neq 0$. This in turn implies that $\pi_5(R_5) = \pi_{5,o}(R_5, R_4)$. But $\pi_5(R_5)$ contains the essential element $\alpha_5 = \{C_5\}$, and since πC_5 has degree two, it follows from Theorem 8 that $\pi_5(R_5)$ is the free cyclic group generated by α_5 . Since $\alpha_5 = 0$ in R_6 , we have finally that $\pi_5(R_6) = \pi_5(R_n) = 0$ ($n \geq 6$).

9. Continuous vector fields over spheres

The above results will now be applied to the study of continuous vector fields over S^n . Geometrically, a continuous vector field may be thought of as a set of functions $V^i(x)$ ($i = 1, \dots, n+1$; $x \in S^n$) defining a unit vector tangent to S^n at the point x , the functions $V^i(x)$ being continuous over all of S^n . A set of p such fields are said to be *independent* if the vectors of the fields at each point of S^n are independent vectors.

Let P_k denote the point of S^n whose coördinates are δ_{ik} ($i, k = 1, \dots, n+1$). R_{n-p} denotes the subgroup of R_n consisting of all rotations leaving P_k fixed ($k = n-p+2, \dots, n+1$). The coset space R_n/R_{n-p} may be thought of as the space of all sets of p mutually orthogonal points of S^n ; for two elements of R_n are in the same coset of R_{n-p} if and only if their last p columns are identical. These columns define the orthogonal p -tuple associated with the given coset. Conversely, given a set of orthogonal points Q_1, Q_2, \dots, Q_p , the required coset is the set of all rotations carrying P_{n-p+2} into Q_1, \dots , and P_{n+1} into Q_p .

Since $R_{n-p} \subset R_{n-1}$, each coset of R_{n-p} lies in some coset of R_{n-1} . Thus a natural mapping $R_n/R_{n-p} \rightarrow R_n/R_{n-1} = S^n$ is defined; each coset of R_{n-p} is mapped into the coset of R_{n-1} containing it. We refer to this map as the *projection*; it is easily verified that it is a fibre map. If we regard an element of R_n/R_{n-p} as being determined by a set of p mutually orthogonal points, the projection is the p^{th} point of the set.

Since the usual process of orthogonalizing a set of independent vectors can be carried out here, there is no loss of generality in assuming that the vectors of the fields we are dealing with are orthogonal at each point and of unit length.

THEOREM 10. *Every set of p orthogonal vector fields over S^n defines a mapping of S^n into R_n/R_{n-p-1} whose projection into S^n is the identity; conversely, any such map defines a set of p orthogonal vector fields.*

By translating the vectors at any point x to the center of S^n and adjoining the point x , we obtain an orthogonal $(p+1)$ -tuple whose projection is x . This process is continuous and gives the required map. Conversely, given such a map, we define the p orthogonal vectors at x as follows: the given map associates with x an orthogonal $(p+1)$ -tuple, the $(p+1)^{\text{st}}$ point of which is x . Transla-

tion of radius vectors drawn to the first p points to the point x gives the set of vectors at x .

COROLLARY. *There exists a set of p independent vector fields over S^n if and only if the canonical map of S^{n-1} into R_{n-1} is contractible in R_{n-1} into R_{n-p-1} .*

It follows from Theorem 7 that there is no vector field over S^n if n is even. However, if n is odd, it follows from the Lemma used in the proof of Theorem 8 that there is at least one. Over S^3 and S^7 are three and seven orthogonal fields, respectively. These are defined by the mappings $S^3 \rightarrow R_3 \rightarrow S^3$ and $S^7 \rightarrow R_7 \rightarrow S^7$ of degree one given by the matrices B and B obtained in Section 5.

THEOREM 11. *If $n \equiv 3 \pmod{4}$ there at least three orthogonal vector fields over S^n ; if $n \equiv 7 \pmod{8}$ there are at least seven.*

We shall prove the theorem for $n = 4m + 3$; the proof for $n = 8m + 7$ is entirely analogous. Let $V^i(x)$ ($i = 1, 2, 3$; $x \in S^n$) define a set of orthogonal vector fields over S^n ; the components of the vector $V^i(x)$ will be denoted by $V_j^i(x)$ ($j = 1, 2, 3, 4$). We extend $V^i(x)$ to be defined over all of E^4 as follows: if $y \in E^4$, let x be the central projection of y on S^3 , r the distance of y from the origin. Then $V^i(y) = rV^i(x)$ defines a set of three orthogonal vectors at each point of E^4 , and these vectors vanish only at the origin.

If z is a point of E^{4m+4} with coordinates (z_1, \dots, z_{4m+4}) , let x^j ($j = 1, \dots, m+1$) denote the point of E^4 with coordinates $(z_{4j-3}, z_{4j-2}, z_{4j-1}, z_{4j})$. Then we define three vector fields $W^i(z)$ over E^{4m+4} as follows:

$$W_{4j+k}^i(z) = V_k^i(x^{j+1}) \quad (i = 1, 2, 3; j = 0, \dots, m; k = 1, 2, 3, 4).$$

Evidently the vectors $W^i(z)$ are mutually orthogonal at each point z , and if $z \in S^{4m+3}$ they are of unit length. Thus the theorem is proved.

The problem of determining the maximum number of independent fields which can exist over S^n has not yet been solved for general odd n . For $n \equiv 1 \pmod{4}$ the solution appears in

THEOREM 12. *If $n \equiv 1 \pmod{4}$ any two vector fields over S^n are somewhere dependent.*

The theorem is evidently true for $n = 1$. Suppose $n > 1$ and suppose that the theorem were not true. Then by the Corollary to Theorem 10, the canonical map C_{n-1} would be contractible in R_{n-1} into R_{n-3} . Hence the map $G_{n-1}(S^{n-1}) \subset R_{n-2}$ would be homotopic in R_{n-1} to a map $F(S^{n-1}) \subset R_{n-3}$. Hence $\{G_{n-1}\} - \{F\} \in \pi_{n-1,0}(R_{n-2})$. But $\pi[\{G_{n-1}\} - \{F\}] = \pi\{G_{n-1}\} - \pi\{F\} = \pi\{G_{n-1}\} \neq 0$. In the proof of Theorem 8 we have shown that this is impossible.

COROLLARY. *If $n > 1$ and $n \equiv 1 \pmod{4}$ the tangent sphere-bundle of S^n is not simple ([7], p. 788).*

UNIVERSITY OF CHICAGO

BIBLIOGRAPHY

1. W. HUREWICZ, *Beiträge zur Topologie der Deformationen*, Proc. Kon. Akad. van Wetenschappen te Amsterdam, 38 (1935), pp. 112-119, 521-528; 39 (1936), pp. 117-125, 215-224.

2. W. HUREWICZ AND N. E. STEENROD, *Homotopy relations in fibre spaces*, Proc. Nat. Acad. of Sci., 27 (1941), pp. 61-64.
3. H. FREUDENTHAL, *Über die Klassen der Sphärenabbildungen*, Comp. Math., 5 (1937), pp. 299-314.
4. H. HOPF, *Über die Abbildungen der dreidimensionalen Sphäre auf die Kugelfläche*, Math. Ann., 104 (1931), pp. 637-665.
5. L. PONTRJAGIN, *A classification of continuous transformations of a complex into a sphere*, C. R., Acad. des Sc. de l'URSS, 19 (1938), pp. 361-363.
6. H. WHITNEY, *On the theory of sphere-bundles*, Proc. Nat. Acad. of Sci., 26 (1940), pp. 148-153.
7. H. WHITNEY, *Topological properties of differentiable manifolds*, Bull. Am. Math. Soc., 43 (1937), pp. 785-805.
8. E. CARTAN, *La topologie des groupes de Lie*, Actualités Scientifiques et Industrielles, Exposés de géométrie, vol. VIII. Paris, 1936.

ON MATRIX ALGEBRAS OVER AN ALGEBRAICALLY CLOSED FIELD*†

BY WINSTON M. SCOTT

(Received June 2, 1941)

Introduction

Recently a number of writers have discussed interesting developments in the theory of not completely reducible matrix sets and non-semisimple algebras.¹ Here we have made use of some of these concepts and methods to study matrix algebras over an algebraically closed field. We shall discuss in some detail in this introduction the definitions that are used and the theorems that are developed.

Let \mathfrak{A} denote a matrix algebra, with unit element, over an algebraically closed field K . \mathfrak{A} will be taken in reduced form, by which we shall mean that \mathfrak{A} is exhibited with only zeros above the main diagonal, with irreducible constituents of \mathfrak{A} in the main diagonal, and that \mathfrak{A} is expressible as a direct sum of its radical and a semisimple subalgebra which latter has non-zero components only in the irreducible constituents of \mathfrak{A} .²

$$(1) \quad \mathfrak{A} = \begin{pmatrix} \mathfrak{C}_{11} & & & \\ \mathfrak{C}_{21} & \mathfrak{C}_{22} & & \\ \vdots & & \ddots & \\ \mathfrak{C}_{i1} & \mathfrak{C}_{i2} & \cdots & \mathfrak{C}_{ii} \end{pmatrix},$$

the \mathfrak{C}_{ii} denoting the irreducible constituents; further

$$\mathfrak{A} = \mathfrak{N} + \mathfrak{A}^*,$$

* The following constitutes a portion of a dissertation written under the direction of Dr. C. J. Nesbitt and accepted by the University of Michigan in January, 1941. The aid and encouragement given me by Dr. Nesbitt has been invaluable in the preparation of this work.

† Presented to the Society, May 2, 1941.

¹ See, for instance, the forthcoming paper by R. Brauer, *On Sets of Matrices with Coefficients in a Division Ring*; T. Nakayama, *On Frobeniusean Algebras*, I, these Annals, Vol. 40 (1939), pp. 611-633; T. Nakayama, *Some Studies on Regular Representations, Induced Representations, and Modular Representations*, these Annals, Vol. 39 (1938), pp. 361-369; T. Nakayama and C. Nesbitt, *Note on Symmetric Algebras*, these Annals, Vol. 39 (1938), pp. 659-668; R. Brauer and C. Nesbitt, *On the Regular Representations of Algebras*, Proc. Nat. Acad. of Sci., Vol. 23 (1937), pp. 236-240; C. Nesbitt, *On the Regular Representations of Algebras*, these Annals, Vol. 39 (1938), pp. 634-658. We shall refer to the last two as B.N. and N.R., respectively.

² Cf. N.R., p. 639.

where \mathfrak{N} is the radical of \mathfrak{A} and

$$(2) \quad \mathfrak{N} = \begin{pmatrix} 0 & & & \\ \mathfrak{C}_{21} & 0 & & \\ \vdots & & \ddots & \\ \mathfrak{C}_{n1} & \mathfrak{C}_{n2} & \dots & 0 \end{pmatrix}, \quad \mathfrak{A}^* = \begin{pmatrix} \mathfrak{C}_{11} & & & \\ 0 & \mathfrak{C}_{22} & & \\ \vdots & & \ddots & \\ 0 & \dots & \dots & \mathfrak{C}_{nn} \end{pmatrix}.$$

Let F_1, F_2, \dots, F_k denote the totality of distinct irreducible representations of \mathfrak{A} . Then each \mathfrak{C}_{ii} is equivalent to one of the F_k . It is well known that up to equivalence the irreducible constituents \mathfrak{C}_{ii} of \mathfrak{A} are uniquely determined.

As a part of \mathfrak{A} , \mathfrak{C}_{ij} forms an additive group or module of matrices upon which \mathfrak{A} , itself considered as a module, is homomorphically mapped. We wish to consider \mathfrak{C}_{ij} as a matrix module with \mathfrak{A} as both a left and right operator system. For a matrix A of \mathfrak{A} , let us use the notation $C_{ij}(A)$, ($j \leq i, i = 1, 2, \dots, t$) to denote the parts of A ,

$$(3) \quad A = \begin{pmatrix} C_{11}(A) & & & \\ C_{21}(A) & C_{22}(A) & & \\ \vdots & \vdots & \ddots & \\ C_{n1}(A) & C_{n2}(A) & \dots & C_{nn}(A) \end{pmatrix}.$$

Let B be any element of \mathfrak{A} , and let B^* be the component of B in the semisimple subalgebra \mathfrak{A}^* . We define B as a left and as a right operator of $C_{ij}(A)$ by the relations below, using \circ to distinguish this operation from ordinary matrix multiplication

$$(4) \quad \begin{aligned} B \circ C_{ij}(A) &= C_{ii}(B) \cdot C_{ij}(A) = C_{ij}(B^*A), \\ C_{ij}(A) \circ B &= C_{ij}(A) \cdot C_{jj}(B) = C_{ij}(AB^*). \end{aligned}$$

Let us call a module which has \mathfrak{A} as both right and left operator system an $(\mathfrak{A}, \mathfrak{A})$ module, and a homomorphism which is an operator mapping under the system of operators \mathfrak{A} , acting on both sides, an $(\mathfrak{A}, \mathfrak{A})$ homomorphism. Under definition (4), \mathfrak{C}_{ij} is a simple $(\mathfrak{A}, \mathfrak{A})$ module. Moreover, each \mathfrak{C}_{ij} is either a 0-part or there exist elements in \mathfrak{A} such that the corresponding C_{ij} have any arbitrary components from K .

Accordingly, we shall call the \mathfrak{C}_{ij} the *simple parts* of \mathfrak{A} . The simple parts \mathfrak{C}_{ii} , or irreducible constituents of \mathfrak{A} , have been rather thoroughly studied. Our first aim, then, is to develop a theory of simple parts which includes the \mathfrak{C}_{ij} , with $i \neq j$, that is, the simple parts of \mathfrak{A} which appear in the radical \mathfrak{N} of \mathfrak{A} .

A first remark is that, when \mathfrak{A} is considered as an $(\mathfrak{A}, \mathfrak{A})$ module, it can be proved in a direct manner that each \mathfrak{C}_{ij} which is not a 0-part is $(\mathfrak{A}, \mathfrak{A})$ isomorphic to a composition factor group of \mathfrak{A} . This result may also be obtained by applying a useful theorem given recently by R. Brauer.³

³ R. Brauer, op. cit.

When \mathbb{C}_{ii} , \mathbb{C}_{jj} are respectively the irreducible constituents F_κ , F_λ of \mathfrak{A} , we shall say that \mathbb{C}_{ij} is of type (κ, λ) . If $\kappa \neq \lambda$, we shall say that \mathbb{C}_{ij} is of mixed type, and if $\kappa = \lambda$ that \mathbb{C}_{ij} is of unmixed type. An $(\mathfrak{A}, \mathfrak{A})$ isomorphism may be defined for any two \mathbb{C}_{ij} of type (κ, λ) and, as a consequence, the corresponding factor groups of \mathfrak{A} are also isomorphic. It is easy to see that two simple parts not of the same type are not $(\mathfrak{A}, \mathfrak{A})$ isomorphic. A representation \mathfrak{A} of \mathfrak{A} may be considered as an $(\mathfrak{A}, \mathfrak{A})$ module. Again applying Brauer's Theorem it follows that the simple parts of type (κ, λ) of a representation \mathfrak{A} , which is in reduced form, are $(\mathfrak{A}, \mathfrak{A})$ isomorphic to the simple parts of \mathfrak{A} of type (κ, λ) .

In N.R. extensive use was made of a basis system of \mathfrak{A} first employed by Cartan⁴, and which was there called the Cartan basis system. The coefficients in the expressions for the elements A of \mathfrak{A} as linear combinations of the Cartan basis elements yield a set of matrix modules \mathfrak{S}^μ , ($\mu = 1, 2, \dots, m$) upon which \mathfrak{A} , as a module, is homomorphically mapped. The \mathfrak{S}^μ may also be considered as $(\mathfrak{A}, \mathfrak{A})$ modules. We shall call these \mathfrak{S}^μ *elementary (related) modules* of \mathfrak{A} . The number m of elementary modules is equal to the composition length of \mathfrak{A} considered as an $(\mathfrak{A}, \mathfrak{A})$ module, and each \mathfrak{S}^μ is $(\mathfrak{A}, \mathfrak{A})$ isomorphic to a composition factor group of \mathfrak{A} . Conversely, to each composition factor group of a composition series of \mathfrak{A} there corresponds an isomorphic elementary module. It follows that up to $(\mathfrak{A}, \mathfrak{A})$ isomorphism the set of elementary modules is uniquely determined.

We say that an elementary module \mathfrak{S}^μ is of type (κ, λ) if \mathfrak{S}^μ is defined by Cartan basis elements of type (κ, λ) . Simple parts and elementary modules may also be classified by the powers of the radical to which they belong. A simple part \mathbb{C}_{ij} (an elementary module \mathfrak{S}^μ) *belongs to* \mathfrak{N}^h if this is the highest power of the radical such that all its elements are not mapped on the 0-matrix in \mathbb{C}_{ij} (\mathfrak{S}^μ). The principal relation between simple parts and elementary modules is that every simple part \mathbb{C}_{ij} of type (κ, λ) which belongs to \mathfrak{N}^h is expressible as a linear combination of elementary modules \mathfrak{S}^μ of type (κ, λ) which belong to \mathfrak{N}^j , $j \leq h$.

The form of a center element of \mathfrak{A} can be computed. For a center element Z , $C_{ij}(Z) = 0$ if \mathbb{C}_{ij} is a simple part of mixed type, and $C_{ij}(Z)$ is of form $(c \cdot \delta_{mn})$, c and element of K , if \mathbb{C}_{ij} is of unmixed type. Let c_α denote, as in B.N., the Cartan Invariants. Then the rank ρ of the center is at most equal to the number $s = \sum_{\kappa=1}^k c_\kappa$ of elementary modules of unmixed type.

As an application of these concepts we obtain a decomposition of the matrix algebra \mathfrak{A} . By use of the concept of type, the simple parts of \mathfrak{A} may be classified into 'blocks'. If \mathbb{C}_{ii} , \mathbb{C}_{jj} , $j < i$, do not belong to the same block then \mathbb{C}_{ij} must be a 0-part. An immediate consequence is that \mathfrak{A} may be decomposed into parts which contain all simple parts belonging to a block. Each such part defines an invariant subalgebra which may not be decomposed into a direct sum of invariant subalgebras.

⁴ E. Cartan, *Les groupes bilinéaires et les systèmes de nombres complexes*, Annales de Toulouse, 12B (1898), p. 1.

1

We consider a matrix algebra \mathfrak{A} , with unit element E , over an algebraically closed field K . We may assume that \mathfrak{A} has no 0-constituent.⁵ We take \mathfrak{A} in the reduced form (1) and define operators for the parts \mathfrak{C}_{ij} of \mathfrak{A} by (4). The field K may be considered as included in the operator system by identifying the element k of K with the element $k \cdot E$ of \mathfrak{A} .

By means of (4) the parts \mathfrak{C}_{ij} of \mathfrak{A} (cf. (1)) may be regarded as $(\mathfrak{A}, \mathfrak{A})$ modules. If \mathfrak{C}_{ij} contains only the 0-matrix, we call \mathfrak{C}_{ij} a 0-part, and obviously, in this case, \mathfrak{C}_{ij} is a simple module. Suppose, instead, that there exists an element A of \mathfrak{A} such that $C_{ij}(A) \neq 0$. It is well known that elements L of \mathfrak{A} can be chosen so that the corresponding $C_{ii}(L)$, $C_{jj}(L)$ have any components arbitrarily chosen from K . It follows from (4) that by suitably choosing right and left operators L , M , \dots to be applied to $C_{ij}(A)$ one obtains a system of matrices

$$C_{ij}(L^* A M^*) = C_{ii}(L) C_{ij}(A) C_{jj}(M)$$

in \mathfrak{C}_{ij} , from which may be generated matrices of \mathfrak{C}_{ij} with arbitrary coefficients. Thus, the admissible subgroup which contains $C_{ij}(A)$ is the whole module \mathfrak{C}_{ij} . Let us call a module of matrices, each of m rows and n columns and with coefficients in K , a *complete module* if it contains all such matrices. We then have

THEOREM 1: *Each \mathfrak{C}_{ij} is either a 0-part or a complete module. In either case \mathfrak{C}_{ij} is a simple $(\mathfrak{A}, \mathfrak{A})$ module.*

We shall let \mathfrak{N}^r denote the set which contains all elements of the form $\sum N_{\mu_1} N_{\mu_2} \cdots N_{\mu_r}$, where the N_{μ_i} are elements contained in the radical \mathfrak{N} . For a sufficiently large value of t , $\mathfrak{N}^t = 0$. These \mathfrak{N}^r form invariant subalgebras of \mathfrak{A} . If we denote \mathfrak{A} by \mathfrak{N}^0 , we have

$$\mathfrak{A} = \mathfrak{N}^0 \supset \mathfrak{N}^1 \supset \mathfrak{N}^2 \supset \mathfrak{N}^3 \supset \cdots \supset \mathfrak{N}^{t-1} \supset \mathfrak{N}^t = 0.$$

We say that an element A belongs to \mathfrak{N}^h if h is the largest value such that \mathfrak{N}^h contains A . In particular, the elements of \mathfrak{A}^* belong to \mathfrak{N}^0 . The product of an element belonging to \mathfrak{N}^h with an element belonging to \mathfrak{N}^j belongs to \mathfrak{N}^k with $k \geq h + j$.

We can now prove

THEOREM 2: *A non-zero simple part \mathfrak{C}_{ij} of \mathfrak{A} is $(\mathfrak{A}, \mathfrak{A})$ isomorphic to a composition factor group of \mathfrak{A} itself considered as an $(\mathfrak{A}, \mathfrak{A})$ module.*

PROOF: Take the series

$$\mathfrak{A} \supset \mathfrak{N}^1 \supset \mathfrak{N}^2 \supset \mathfrak{N}^3 \supset \cdots \supset \mathfrak{N}^{t-1} \supset \mathfrak{N}^t = 0$$

and refine it to obtain a composition series

$$\mathfrak{A} \supset \cdots \supset \mathfrak{A}_{q-1} \supset \mathfrak{A}_q \supset \cdots \supset \mathfrak{A}_{n-1} \supset 0$$

⁵ For, from conditions stated for \mathfrak{A} , we may show that if \mathfrak{A} has 0-constituents, then it may be decomposed in the form $\mathfrak{A} = \begin{pmatrix} \mathfrak{A}_1 & 0 \\ 0 & 0 \end{pmatrix}$, where \mathfrak{A}_1 does not have 0-constituents, and \mathfrak{A} is isomorphic to \mathfrak{A}_1 .

of \mathfrak{A} . Let \mathfrak{A}_q be the first group in this series such that for each of its elements A , $C_{ij}(A) = 0$. Then denoting elements of \mathfrak{A}_{q-1} by A_{q-1} , we have that

$$A_{q-1} \rightarrow C_{ij}(A_{q-1})$$

is an $(\mathfrak{A}, \mathfrak{A})$ homomorphism, since for any element of \mathfrak{A} ,

$$\begin{aligned} B \cdot A_{q-1} &= B^* \cdot A_{q-1} + N \cdot A_{q-1} \\ &\rightarrow C_{ij}(B^* A_{q-1}) + C_{ij}(N \cdot A_{q-1}). \end{aligned}$$

But $N \cdot A_{q-1} \subset \mathfrak{A}_q$ so that $C_{ij}(N A_{q-1}) = 0$ and, therefore,

$$B \cdot A_{q-1} \rightarrow C_{ij}(B^* A_{q-1}) = B \circ C_{ij}(A_{q-1}).$$

The same argument holds for right side multiplication. The kernel of this mapping (that is, the elements mapping into the identity element) contains \mathfrak{A}_q and, therefore, the kernel is \mathfrak{A}_q . It follows that $\mathfrak{A}_{q-1}/\mathfrak{A}_q$ is $(\mathfrak{A}, \mathfrak{A})$ isomorphic to the simple part \mathfrak{S}_{ij} .

We have said that if \mathfrak{S}_{ii} , \mathfrak{S}_{jj} are respectively the irreducible constituents F_κ , F_λ of \mathfrak{A} , then the simple part \mathfrak{S}_{ij} is of type (κ, λ) . We now show that

THEOREM 3: *The non-zero simple parts \mathfrak{S}_{ij} of a given type (κ, λ) are mutually isomorphic in the $(\mathfrak{A}, \mathfrak{A})$ sense. Simple parts of different types are not isomorphic.*

PROOF: It should be noted here that the word *isomorphic* is to have a meaning differing from that of Theorem 2. In Theorem 2, the isomorphism is obtained from $A \rightarrow C_{ij}(A)$, that is, \mathfrak{S}_{ij} is considered as a set of matrices related to \mathfrak{A} . Here, \mathfrak{S}_{ij} is just to represent the system of matrices.

Let \mathfrak{S}_{ij} , \mathfrak{S}_{mn} be two simple parts of type (κ, λ) which are not 0-parts. An $(\mathfrak{A}, \mathfrak{A})$ mapping of \mathfrak{S}_{ij} upon \mathfrak{S}_{mn} is obtained by the relation $C_{ij} \rightarrow C_{mn}$ if $C_{ij} = C_{mn}$ where C_{ij} , C_{mn} denote matrices of \mathfrak{S}_{ij} , \mathfrak{S}_{mn} , respectively. For if B is any element of \mathfrak{A} , then

$$\begin{aligned} B \circ C_{ij} &= C_{ii}(B) \cdot C_{ij} = F_\kappa(B) \cdot C_{ij} \\ &= C_{mn}(B) \cdot C_{mn} = B \circ C_{mn}. \end{aligned}$$

On the other hand, if \mathfrak{S}_{ij} , \mathfrak{S}_{mn} are of types (κ, λ) , (μ, ν) , respectively, with $\kappa \neq \mu$, then any relation $C_{ij} \rightarrow C_{mn}$ is not an $(\mathfrak{A}, \mathfrak{A})$ operator mapping, for we may choose B in \mathfrak{A} such that $C_{ii}(B) = F_\kappa(B) = 0$ while $C_{mn}(B) = F_\mu(B) = E_{f_\mu}$ where E_{f_μ} is a unit matrix of proper degree.

2

The basis of \mathfrak{A} which seems best adapted to the study of simple parts is a basis system first employed by Cartan.⁶ This basis was used by C. Nesbitt in his study of regular representations of linear associative algebras.⁷

We denote, as before, the irreducible representations of \mathfrak{A} by F_1 , F_2 , F_3 ,

⁶ E. Cartan, op. cit.

⁷ A detailed description of the Cartan basis system is given in N.R. section 2.

\dots, F_k , and their degrees by f_1, f_2, \dots, f_k . Let $E_\kappa(m, n)$ be the matrix of \mathfrak{A} which has 1 for the (m, n) component of each \mathfrak{C}_i which is equal to F_κ , and has zero elsewhere. The relations

$$E_\mu(m, n) \cdot E_\nu(p, q) = \begin{cases} 0 & \text{if either } \mu \neq \nu, \text{ or } p \neq q, \\ E_\mu(m, q) & \text{if } \mu = \nu, n = p, \end{cases}$$

follow from the definition of the $E_\mu(m, n)$.

Let $E_\kappa = \sum_i E_\kappa(ii)$. Then the E_κ , ($\kappa = 1, 2, \dots, k$), form a system of mutually idempotent elements. In fact, E_κ is the unit element of a certain simple algebra which is a summand in a decomposition of \mathfrak{A}^* into a direct sum of simple algebras, and $\sum_{\kappa=1}^k E_\kappa = E$, where E is the unit element of \mathfrak{A} .

A matrix A of \mathfrak{A} is said to be of type (κ, λ) if $E_\kappa \cdot A \cdot E_\lambda = A$. In particular, $E_\kappa(m, n)$ is of type (κ, κ) . An element of type (κ, λ) , $\kappa \neq \lambda$, we shall say is of mixed type, and if $\kappa = \lambda$, of unmixed type. The product of an element of type (κ, λ) and an element of type (μ, ν) is an element of type (κ, ν) . The product of an element of type (κ, λ) and of an element of type (μ, ν) vanishes for $\lambda \neq \mu$. A matrix A of type (κ, λ) has non-zero components only in simple parts \mathfrak{C}_{ij} of type (κ, λ) .

A basis set $B_{\kappa\lambda}^1, B_{\kappa\lambda}^2, \dots, B_{\kappa\lambda}^{c_{\kappa\lambda}}$ is first obtained for those matrices of \mathfrak{A} of type (κ, λ) which have non-zero components only in the upper left corners of simple parts of type (κ, λ) . The rank t of this set is the Cartan Invariant, $c_{\kappa\lambda}$.⁸ A basis for the whole algebra is then given by the set

$$\begin{aligned} E_\kappa(a1)B_{\kappa\lambda}^\mu E_\lambda(1b), \\ \mu = 1, 2, \dots, c_{\kappa\lambda}, \quad a = 1, 2, \dots, f_\kappa, \\ b = 1, 2, \dots, f_\lambda, \quad \kappa, \lambda = 1, 2, \dots, k. \end{aligned}$$

This basis can be so chosen that any element A of \mathfrak{A} which belongs to \mathfrak{N}^τ is expressible in terms of basis elements which belong to \mathfrak{N}^ρ , for $\rho \geq \tau$.⁹

For convenience we change slightly the above notation. We take together in one set all the elements $B_{\kappa\lambda}^\mu$ for all κ, λ , and μ and redefine the superscript μ to enumerate these elements, independently of type, in such a way as to take first all those elements which belong to \mathfrak{N}^0 , namely $E_1(11), \dots, E_k(11)$, then next those which belong to \mathfrak{N}^1 , and so on. μ will now have the range $1, 2, \dots, m$ where $m = \sum_{\kappa, \lambda=1}^k c_{\kappa\lambda}$. Where the type of the element need not be explicitly stated we shall drop the subscripts κ, λ from the symbols $B_{\kappa\lambda}^\mu$.

3

Expressing the element A of \mathfrak{A} as a linear combination of the elements of the Cartan basis system, we have

$$(5) \quad A = \sum_{a,b,\mu} h_{ab}^\mu(A) E_\kappa(a1) B_{\kappa\lambda}^\mu E_\lambda(1b).$$

⁸ See B.N. or N.R.

⁹ See N.R., p. 641.

Here each $B_{\kappa\lambda}^\mu$ is a basis element having components different from zero only in the upper left hand corners of simple parts \mathfrak{S}_{ij} which are of type (κ, λ) .

We denote the matrix $(h_{ab}^\mu(A))_{ab}$ (cf. (5)) by $H^\mu(A)$. The module consisting of the matrices $H^\mu(A)$ will be denoted by \mathfrak{S}^μ . We shall call \mathfrak{S}^μ an *elementary (related) module* of \mathfrak{A} . \mathfrak{S}^μ is evidently a complete module, since the coefficients in (5) may be taken arbitrarily from K .

An elementary module will be said to be of type (κ, λ) if it results from, or is defined by, Cartan basis elements of type (κ, λ) . We shall write $\mathfrak{S}_{\kappa\lambda}^\mu$ to denote that \mathfrak{S}^μ is of type (κ, λ) . If $\kappa \neq \lambda$ we shall call $\mathfrak{S}_{\kappa\lambda}^\mu$ an elementary module of *mixed type*, and if $\kappa = \lambda$ we shall call $\mathfrak{S}_{\kappa\lambda}^\mu$ an elementary module of *unmixed type*.

Let us say that a module \mathfrak{M} of matrices is related to \mathfrak{N}^ρ if \mathfrak{N}^ρ is homomorphically mapped on \mathfrak{M} . In general, we shall deal with modules \mathfrak{M} which are related to $\mathfrak{N}^0 = \mathfrak{A}$, but it is convenient sometimes to consider \mathfrak{M} as the map of only the subalgebra \mathfrak{N}^ρ rather than \mathfrak{A} itself (see p. 157). A related module \mathfrak{M} will be said to belong to \mathfrak{N}^σ if σ is the largest value such that for at least one element N_σ belonging to \mathfrak{N}^σ , the corresponding matrix $M(N_\sigma)$ of \mathfrak{M} is not zero. In particular, the elementary module \mathfrak{S}^μ is related to \mathfrak{A} , and belongs to the same power of \mathfrak{N} as does the corresponding basis element B^μ .

LEMMA 1: Let \mathfrak{M} be a module for which the concept of type is defined. If

- i). \mathfrak{M} is of type (κ, λ) , and its elements are $f_\kappa \times f_\lambda$ matrices,
- ii). \mathfrak{M} is related to \mathfrak{N}^ρ , and belongs to \mathfrak{N}^σ , and
- iii). \mathfrak{M} is a simple $(\mathfrak{A}, \mathfrak{N})$ module such that for any element B of \mathfrak{A} , and element N_ρ of \mathfrak{N}^ρ

$$(6) \quad \begin{aligned} B \circ M(N_\rho) &= F_\kappa(B) \cdot M(N_\rho) = M(B^* N_\rho) \\ M(N_\rho) \circ B &= M(N_\rho) \cdot F_\lambda(B) = M(N_\rho B^*), \end{aligned}$$

then it follows that as a module related to \mathfrak{N}^ρ , $M(N_\rho) = \sum_\mu m_\mu H_{\kappa\lambda}^\mu(N_\rho)$, where m_μ are fixed scalars, N_ρ is any element of \mathfrak{N}^ρ , and where the elementary module $\mathfrak{S}_{\kappa\lambda}^\mu$ belongs to \mathfrak{N}^τ , $\rho \leq \tau \leq \sigma$, and is considered as related to \mathfrak{N}^σ .

PROOF: We shall use the notation $(U_{rs})_{pq}$ to denote a matrix with 1 as its (r, s) component, and with zero elsewhere. Then operating with $E_\kappa(1, 1)$, $E_\lambda(1, 1)$ on $M(B_{\kappa\lambda}^\mu) = (m_{pq}^\mu)$, we obtain

$$\begin{aligned} E_\kappa(11) \circ M(B_{\kappa\lambda}^\mu) \circ E_\lambda(11) &= (U_{11})_{rs} \cdot (m_{pq}^\mu) \cdot (U_{11})_{uv} = (m_{11}^\mu) \cdot (U_{11})_{pq} \\ &= M(E_\kappa(11) B_{\kappa\lambda}^\mu E_\lambda(11)) = M(B_{\kappa\lambda}^\mu). \end{aligned}$$

Dropping the subscripts 11 we have

$$M(B_{\kappa\lambda}^\mu) = (m^\mu U_{11})_{pq}.$$

Here $m^\mu \neq 0$ only if $B_{\kappa\lambda}^\mu$ belongs to \mathfrak{N}^τ , $\tau \leq \sigma$. Also

$$M(E_\kappa(a1) B_{\kappa\lambda}^\mu E_\lambda(1b)) = (m^\mu U_{ab})_{pq}.$$

We have finally that

$$\begin{aligned} M(N_\rho) &= M\left(\sum_{\mu, a, b} h_{ab}^\mu(N_\rho) E_\kappa(a1) B_{\kappa\lambda}^\mu E_\lambda(1b)\right) \\ &= \sum_\mu m^\mu H_{\kappa\lambda}^\mu(N^\rho) \end{aligned}$$

where $\mathfrak{S}_{\kappa\lambda}^\mu$ belongs to \mathfrak{N}^τ , $\rho \leq \tau \leq \sigma$, since N_ρ is an element of \mathfrak{N}^ρ , and m belongs to \mathfrak{N}^σ .

Relating the simple parts of \mathfrak{A} to the elementary modules of \mathfrak{A} , we have

THEOREM 4: *A simple part \mathfrak{S}_{ij} of \mathfrak{A} of type (κ, λ) which belongs to \mathfrak{N}^τ is expressible as a linear combination of the elementary modules of type (κ, λ) which belong to \mathfrak{N}^τ , $\tau \leq r$.*

PROOF: We consider \mathfrak{S}_{ij} as a simple module related to $\mathfrak{A} = \mathfrak{N}^0$, and belonging to \mathfrak{N}^τ , and apply the lemma. It follows that \mathfrak{S}_{ij} is of the form $\sum_\mu d_{ij}^\mu H_{\kappa\lambda}^\mu$, that is, for each element A of \mathfrak{A} we have that

$$C_{ij}(A) = \sum_\mu d_{ij}^\mu H_{\kappa\lambda}^\mu(A), \quad .$$

where each $H_{\kappa\lambda}^\mu(A)$ belongs to some N^τ , $\tau \leq r$.

We now prove

THEOREM 5: *The number m of elementary modules is equal to the composition length of \mathfrak{A} considered as an $(\mathfrak{A}, \mathfrak{A})$ module. Each elementary module \mathfrak{S} is $(\mathfrak{A}, \mathfrak{A})$ isomorphic to a composition factor group of \mathfrak{A} , and conversely, to each factor group of \mathfrak{A} there corresponds an isomorphic elementary module.*

PROOF: We will construct a composition series for \mathfrak{A} in terms of the Cartan basis elements. As indicated above, we take the whole set of elements $B_{\kappa\lambda}^\mu$ ($\kappa, \lambda = 1, 2, 3, \dots, k$) and use the superscript μ ($\mu = 1, 2, 3, \dots, m$) to enumerate these elements, taking first those elements which are not in \mathfrak{N}^1 , then those which are in \mathfrak{N}^1 but not in \mathfrak{N}^2 , and so on. Let \mathfrak{A}_{s-1} denote the subgroup of \mathfrak{A} consisting of all linear combinations with coefficients in K of the elements

$$\begin{aligned} &(\kappa, \lambda = 1, 2, \dots, \text{or } k), \\ (7) \quad &E_\kappa(a1) B_{\kappa\lambda}^\tau E_\lambda(1b), \quad (a = 1, 2, \dots, f_\kappa; \quad b = 1, 2, \dots, f_\lambda), \\ &(\tau = S, S+1, \dots, m). \end{aligned}$$

In particular, $\mathfrak{A}_0 = \mathfrak{A}$. We form the product

$$A \cdot E_\kappa(a1) B_{\kappa\lambda}^\tau E_\lambda(1b) = (A^* + N) \cdot E_\kappa(a1) B_{\kappa\lambda}^\tau E_\lambda(1b).$$

Here

$$\begin{aligned} A^* \cdot E_\kappa(a1) B_{\kappa\lambda}^\tau E_\lambda(1b) &= \sum_{\rho, m, n} f_{mn}^\rho(A) E_\rho(mn) \cdot E_\kappa(a1) B_{\kappa\lambda}^\tau E_\lambda(1b) \\ &= \sum_m f_{ma}^\kappa(A) E_\kappa(m1) B_{\kappa\lambda}^\tau E_\lambda(1b) \end{aligned}$$

is again in \mathfrak{A}_{s-1} . So also is $N \cdot E_\kappa(a1) B_{\kappa\lambda}^\tau E_\lambda(1b)$, since this belongs to a higher

power of the radical than does $B_{\kappa\lambda}^s$ and is, therefore, expressible in terms of the elements (7). Then the whole product is in \mathfrak{A}_{s-1} . The same is true when A is a right factor, so we obtain that \mathfrak{A}_{s-1} is an admissible subgroup of \mathfrak{A} . Moreover, the factor group $\mathfrak{A}_{s-1}/\mathfrak{A}_s$ consisting of the residue classes

$$\langle \sum_{a,b} \lambda_{ab}^s E_{\kappa}(a1) B_{\kappa\lambda}^s E_{\lambda}(1b) \rangle, \quad \text{modulo } \mathfrak{A}_s,$$

is easily seen by similar calculations to be simple. It follows that

$$\mathfrak{A}_0 \supset \mathfrak{A}_1 \supset \mathfrak{A}_2 \supset \cdots \supset \mathfrak{A}_{m-1} \supset 0$$

is a composition series of \mathfrak{A} , and the first part of the theorem is proved.

Again, we may easily calculate that

$$\langle \sum_{a,b} h_{ab}^s(A) E_{\kappa}(a1) B_{\kappa\lambda}^s E_{\lambda}(1b) \rangle \rightarrow H_{\kappa\lambda}^s(A)$$

is an $(\mathfrak{A}, \mathfrak{A})$ isomorphic mapping of $\mathfrak{A}_{s-1}/\mathfrak{A}_s$ upon the elementary module $\mathfrak{S}_{\kappa\lambda}^s$ (cf. (7)). From this follow the other statements in the theorem.

It may be noted here that from Theorem 4 and Theorem 5 another proof of Theorem 2 may be obtained.

We can now prove the converse of Theorem 2

THEOREM 6: *To each composition factor group of \mathfrak{A} there corresponds at least one non-zero simple part \mathfrak{C}_{ij} .*

PROOF: By Theorem 5 we have that to each composition factor group of \mathfrak{A} there corresponds an isomorphic elementary module \mathfrak{S}^p , say of type (μ, ν) . Then there exist non-zero simple parts of type (μ, ν) , and by the same argument as that used in Theorem 3 it may be shown that each non-zero simple part of type (μ, ν) is $(\mathfrak{A}, \mathfrak{A})$ isomorphic to \mathfrak{S}^p .

4

A system of matrices $\bar{\mathfrak{A}}$ is called a representation of \mathfrak{A} if \mathfrak{A} is homomorphically mapped on $\bar{\mathfrak{A}}$. We may define \mathfrak{A} as a left and right operator system for a representation $\bar{\mathfrak{A}}$ by the relations

$$(8) \quad \begin{aligned} A \cdot \bar{A}(A_1) &= \bar{A}(A \cdot A_1) = \bar{A}(A) \cdot \bar{A}(A_1), \\ \bar{A}(A_1) \cdot A &= \bar{A}(A_1 \cdot A) = \bar{A}(A_1) \cdot \bar{A}(A). \end{aligned}$$

Further, if the representation $\bar{\mathfrak{A}}$ is taken in reduced form, then a simple part $\bar{\mathfrak{C}}_{ij}$ of $\bar{\mathfrak{A}}$ may be considered as an $(\mathfrak{A}, \mathfrak{A})$ module if we define the elements of \mathfrak{A} as operators in $\bar{\mathfrak{C}}_{ij}$ by relations of the form (4). We then obtain

THEOREM 7: *The simple parts $\bar{\mathfrak{C}}_{ij}$, of type (κ, λ) , of a representation $\bar{\mathfrak{A}}$ of \mathfrak{A} are $(\mathfrak{A}, \mathfrak{A})$ isomorphic¹⁰ to simple parts \mathfrak{C}_{mn} , of type (κ, λ) , of \mathfrak{A} .*

¹⁰ As in Theorem 3, the word *isomorphism* here has a somewhat different meaning from that of Theorem 2.

PROOF: By Theorem 2 we have that each simple part $\bar{\mathfrak{C}}_{ij}$ of $\bar{\mathfrak{A}}$ is isomorphic to a composition factor group of a composition series for $\bar{\mathfrak{A}}$. Let

$$(9) \quad \bar{\mathfrak{A}} \supset \bar{\mathfrak{A}}_1 \supset \bar{\mathfrak{A}}_2 \supset \dots \supset \bar{\mathfrak{A}}_l = 0,$$

$$(10) \quad \mathfrak{A} \supset \mathfrak{A}_1 \supset \mathfrak{A}_2 \supset \dots \supset \mathfrak{A}_n = 0,$$

be composition series for $\bar{\mathfrak{A}}$, \mathfrak{A} , respectively, and suppose

$$\bar{\mathfrak{C}}_{ij} \cong \bar{\mathfrak{A}}_{p-1}/\bar{\mathfrak{A}}_p.$$

$\bar{\mathfrak{A}}$ is an $(\mathfrak{A}, \mathfrak{A})$ module upon which \mathfrak{A} is homomorphically mapped by the relation $A \rightarrow \bar{A}(A)$. Then the Jordan-Hölder Theorem gives us that to the factor group $\bar{\mathfrak{A}}_{p-1}/\bar{\mathfrak{A}}_p$ of (9) there corresponds a factor group $\mathfrak{A}_{q-1}/\mathfrak{A}_q$ of the series (10) such that

$$\bar{\mathfrak{A}}_{p-1}/\bar{\mathfrak{A}}_p \cong \mathfrak{A}_{q-1}/\mathfrak{A}_q.$$

In addition we have from Theorem 6 that to the factor group $\mathfrak{A}_{q-1}/\mathfrak{A}_q$ there is at least one non-zero isomorphic simple part \mathfrak{C}_{mn} of \mathfrak{A} ,

$$\mathfrak{A}_{q-1}/\mathfrak{A}_q \cong \mathfrak{C}_{mn}.$$

Combining these relations, and observing that all these isomorphisms are $(\mathfrak{A}, \mathfrak{A})$ isomorphisms, we obtain

$$\bar{\mathfrak{C}}_{ij} \cong \mathfrak{C}_{mn}$$

with operators $(\mathfrak{A}, \mathfrak{A})$. This completes the proof of the statement.

Now suppose that $\bar{\mathfrak{A}}$ is a representation of \mathfrak{A} , and that $\bar{\mathfrak{A}}$ is in reduced form. Let us denote the element of $\bar{\mathfrak{A}}$ corresponding to the element A of \mathfrak{A} by \bar{A} , $A \rightarrow \bar{A}$; further, we shall consider $\bar{\mathfrak{A}}$ as a direct sum of its radical $\bar{\mathfrak{A}}_R$ and a semi-simple subalgebra, $\bar{\mathfrak{A}}_S$, $\bar{\mathfrak{A}} = \bar{\mathfrak{A}}_S + \bar{\mathfrak{A}}_R$, where $\bar{\mathfrak{A}}_S$ is completely decomposed into its irreducible constituents. We use \bar{A}_S , \bar{A}_R to denote the components in $\bar{\mathfrak{A}}_S$, $\bar{\mathfrak{A}}_R$ of the element \bar{A} of $\bar{\mathfrak{A}}$, $\bar{A} = \bar{A}_S + \bar{A}_R$.

The set $\bar{\mathfrak{A}}^*$ of elements in the representation $\bar{\mathfrak{A}}$ which correspond to the semi-simple subalgebra \mathfrak{A}^* of \mathfrak{A} form a semisimple subalgebra of $\bar{\mathfrak{A}}$ equivalent to $\bar{\mathfrak{A}}_S$. Since $\bar{\mathfrak{A}}^*$, $\bar{\mathfrak{A}}_S$ have the same irreducible constituents, they are equivalent algebras, but are not necessarily identical. It is easy to construct examples where $\bar{\mathfrak{A}}^*$ is not $\bar{\mathfrak{A}}_S$.

Let us suppose $\bar{\mathfrak{F}}_1, \dots, \bar{\mathfrak{F}}_r$ are a set of elementary modules for the algebra $\bar{\mathfrak{A}}$. We wish to give explicitly the relation between the modules $\bar{\mathfrak{F}}$ of $\bar{\mathfrak{A}}$, and the elementary modules \mathfrak{F} of \mathfrak{A} .

The irreducible constituents of $\bar{\mathfrak{A}}$ are equivalent to irreducible constituents of \mathfrak{A} . For simplicity, let us suppose they are identical with constituents of \mathfrak{A} . Let $\bar{\mathfrak{F}}$ be of type (κ, λ) and belong to the σ^{th} power of the radical of $\bar{\mathfrak{A}}$. We may consider $\bar{\mathfrak{F}}$ as a module related to \mathfrak{A} by setting $\bar{H}(A) = \bar{H}(\bar{A})$. As such, $\bar{\mathfrak{F}}$ belongs to \mathfrak{N}^σ . Let B be any element of \mathfrak{A} , and N_σ an element of \mathfrak{N}^σ . Then

$$F_\kappa(B)\bar{H}(N_\sigma) = F_\kappa(B)\bar{H}(\bar{N}_\sigma) = \bar{H}(B_S\bar{N}_\sigma).$$

But $\bar{B}_s \equiv \bar{B}^*, \text{ mod } \bar{\mathfrak{A}}_R$, so that $\bar{B}_s \bar{N}_\sigma \equiv \bar{B}^* \bar{N}_\sigma, \text{ mod } \bar{\mathfrak{A}}_R^{\sigma+1}$, and as $\bar{\mathfrak{S}}$ belongs to the σ^{th} power of the radical $\bar{\mathfrak{A}}_R$ of $\bar{\mathfrak{A}}$, we have

$$\bar{H}(\bar{B}_s \bar{N}_\sigma) = \bar{H}(\bar{B}^* \bar{N}_\sigma) = \bar{H}(\overline{B^* N_\sigma}) = \bar{H}(B^* N_\sigma).$$

We may then view $\bar{\mathfrak{S}}$ as an $(\bar{\mathfrak{A}}, \bar{\mathfrak{A}})$ module of type (κ, λ) , related and belonging to $\bar{\mathfrak{N}}^\sigma$, such that

$$B \circ \bar{H}(N_\sigma) = F_*(B) \bar{H}(N_\sigma) = \bar{H}(B^* N_\sigma)$$

with a similar relation for right operators. Applying Lemma 1, we have that as a module related to $\bar{\mathfrak{N}}^\sigma$, $\bar{\mathfrak{S}}$ is a linear combination of the elementary modules $\bar{\mathfrak{S}}_{\kappa\lambda}^\mu$ of type (κ, λ) which belong to $\bar{\mathfrak{N}}^\sigma$.¹¹ It follows, also, that a simple part $\bar{\mathfrak{C}}$ of $\bar{\mathfrak{A}}$ of type (κ, λ) which belongs to the σ^{th} power of the radical of $\bar{\mathfrak{A}}$ may, considered as a module related to $\bar{\mathfrak{N}}^\sigma$, be expressed as a linear combination of the elementary modules $\bar{\mathfrak{S}}_{\kappa\lambda}^\mu$ of type (κ, λ) which belong to $\bar{\mathfrak{N}}^\sigma$.

If $\bar{\mathfrak{A}}^* = \bar{\mathfrak{A}}_s$ then the relation

$$F_*(B) \bar{H}(A) = \bar{H}(B^* A)$$

holds for all elements A of $\bar{\mathfrak{A}}$. Again applying Lemma 1 we obtain that $\bar{\mathfrak{S}}$ considered as a module related to $\bar{\mathfrak{A}}$ itself is expressible as a linear combination of the elementary modules $\bar{\mathfrak{S}}$ of type (κ, λ) which belong to $\bar{\mathfrak{N}}^\tau, \tau \leq \sigma$.

If $\bar{\mathfrak{A}}^*$ is not identical to $\bar{\mathfrak{A}}_s$, that is, if $\bar{\mathfrak{A}}^*$ is not completely decomposed, then we can go over to an equivalent representation $\bar{\mathfrak{A}}$, such that $\bar{\mathfrak{A}}^* = \bar{\mathfrak{A}}_s$ where $\bar{\mathfrak{A}}^*, \bar{\mathfrak{A}}_s$ have the same meaning with regard to $\bar{\mathfrak{A}}$ as do $\bar{\mathfrak{A}}^*$ and $\bar{\mathfrak{A}}_s$ in regard to $\bar{\mathfrak{A}}$. We may then state:

THEOREM 8: *To any representation of $\bar{\mathfrak{A}}$ there exists an equivalent representation whose simple parts, considered as modules related to $\bar{\mathfrak{A}}$, are expressible as linear combinations of the elementary modules $\bar{\mathfrak{S}}$ of $\bar{\mathfrak{A}}$.*

5

It is easy now to discuss the center of $\bar{\mathfrak{A}}$. We note that

THEOREM 9: *For a center element Z of $\bar{\mathfrak{A}}$, $C_{ij}(Z) = 0$ if \mathfrak{C}_{ij} is a simple part of mixed type, and $C_{ij}(Z)$ is of form $(c \cdot \delta_{mn})$, c an element of K , if \mathfrak{C}_{ij} is of unmixed type.*

PROOF: Since an element Z of the center must commute with the elements $E_\kappa(mn)$ of $\bar{\mathfrak{A}}$ ($\kappa = 1, 2, \dots, k; m, n = 1, 2, \dots, f_\kappa$) an easy calculation shows that in terms of the Cartan basis elements

$$Z = \sum_{r,r'} c'(Z) E_\kappa(r1) B_{r\kappa}' E_\kappa(1r),$$

¹¹ If we do not take the constituents of $\bar{\mathfrak{A}}$ identical with constituents of $\bar{\mathfrak{A}}$ then it is necessary to apply suitable left and right non-singular matrix multipliers to the $\bar{\mathfrak{S}}_{\kappa\lambda}^\mu$ in the expression for $\bar{\mathfrak{S}}$.

in which the range for τ is determined by the s basis elements B^τ of unmixed type. It follows that $H^\mu(Z) = 0$ if \mathfrak{S}^μ is of mixed type, and

$$H^\mu(Z) = (c^\mu(Z)\delta_{ij})_{ij}$$

if \mathfrak{S}^μ is of unmixed type. Applying Theorem 4 we obtain the theorem.

It should be noted that this theorem can be obtained directly from Schur's Lemma and (2).

6

If an element is of type (κ, λ) we say that κ, λ are the type indices of the element. We say that two non-zero simple parts $\mathfrak{C}_{ij}, \mathfrak{C}_{pq}$, belong to the same block¹² if there exists a chain of non-zero parts such that any two neighboring parts have at least one type index in common. This relation is reflexive, symmetric, and transitive so that a separation of the non-zero simple parts into distinct disjoint classes is obtained.

THEOREM 10: *Each simple part in the i^{th} row of \mathfrak{A} is either a 0-part or belongs to the same block as \mathfrak{C}_{ii} . Similarly, each simple part in the j^{th} column is either a 0-part or belongs to the same block as \mathfrak{C}_{jj} . If $\mathfrak{C}_{ii}, \mathfrak{C}_{jj}$ do not belong to the same block, then \mathfrak{C}_{ij} is a 0-part.*

For if the simple part \mathfrak{C}_{ij} is not a 0-part, then $\mathfrak{C}_{ii}, \mathfrak{C}_{ij}, \mathfrak{C}_{jj}$ is a proper chain and so $\mathfrak{C}_{ii}, \mathfrak{C}_{ij}, \mathfrak{C}_{jj}$ all belong to the same block.

In the same way we may classify the basis elements

$$(11) \quad B_{\kappa\lambda}^\mu, \quad (\mu = 1, 2, \dots, m; \kappa, \lambda = 1, 2, \dots, \text{or } k)$$

into blocks. We say that $B_{\kappa\lambda}^\mu, B_{\rho\sigma}^\gamma$ belong to the same block if there exists a chain of elements $B_{\alpha\beta}^\tau$ of the set (11) connecting $B_{\kappa\lambda}^\mu, B_{\rho\sigma}^\gamma$ such that any two neighboring elements in this chain have at least one type index in common. Let us denote by \mathfrak{B} the set consisting of all the $B_{\rho\sigma}^\tau$ belonging to a given block and their associated elements of form $E_\rho(a1)B_{\rho\sigma}^\tau E_\sigma(1b)$. If A is any element of \mathfrak{A} , then $A \cdot E_\rho(a1)B_{\rho\sigma}^\tau E_\sigma(1b)$ is expressible in terms of \mathfrak{B} again; the similar statement is true for the product $E_\rho(a1)B_{\rho\sigma}^\tau E_\sigma(1b) \cdot A$. It follows that the elements of \mathfrak{B} form a basis of an invariant subalgebra of \mathfrak{A} : we denote this invariant subalgebra by \mathfrak{A}^+ .

We can now prove

THEOREM 11: *By elementary transformations \mathfrak{A} may be decomposed into the form*

$$(12) \quad \mathfrak{A} = \begin{pmatrix} \mathfrak{B}_1 & & 0 \\ & \mathfrak{B}_2 & \\ 0 & & \ddots \\ & & & \mathfrak{B}_i \end{pmatrix}$$

where each \mathfrak{B}_i contains all simple parts belonging to one block. The elements of \mathfrak{A} which have only zeros in the parts $\mathfrak{B}_i, j \neq i$, form an invariant subalgebra \mathfrak{A}_i of \mathfrak{A} ,

¹² R. Brauer and C. Nesbitt, *On the Modular Representations of Finite Groups*, Univ. of Toronto Studies, Math. Series, No. 4 (1937); R. Brauer and C. Nesbitt, *On the Modular Character of Groups*, these Annals, Vol. 42 (1941), pp. 556-590; T. Nakayama, *Some Studies on Regular Representations, Induced Representations, and Modular Representations*, these Annals, Vol. 39 (1938), Theorem 5.

and \mathfrak{A} is the direct sum of these invariant subalgebras \mathfrak{A}_i . The \mathfrak{A}_i may not be decomposed into a direct sum of invariant subalgebras.

PROOF: If \mathfrak{C}_{ii} , $\mathfrak{C}_{i+1,i+1}$ do not belong to the same block, then by Theorem 10, $\mathfrak{C}_{i+1,1}$ is a 0-part. It follows that by permuting the i^{th} row with the $(i+1)^{\text{st}}$ row and the i^{th} column with the $(i+1)^{\text{st}}$ column that we may reverse the positions of \mathfrak{C}_{ii} , $\mathfrak{C}_{i+1,i+1}$ in the main diagonal of \mathfrak{A} , but leave \mathfrak{A} in reduced form. By successive transformations of this form we may bring together in an upper part \mathfrak{B}_1 of \mathfrak{A} all \mathfrak{C}_{ii} which belong to the same block as \mathfrak{C}_{11} . We then have

$$\mathfrak{A} = \begin{pmatrix} \mathfrak{B}_1 & 0 \\ \mathfrak{D}_{21} & \mathfrak{D}_{22} \end{pmatrix}.$$

But since the simple parts \mathfrak{C}_{ij} in \mathfrak{B}_1 and the \mathfrak{C}_{ii} in \mathfrak{D}_{22} belong now to different blocks, then the simple parts in \mathfrak{D}_{21} must all be 0-parts, or, that is, \mathfrak{D}_{21} is a 0-part. By continuing this process we obtain \mathfrak{A} in the form (12).

To prove the remaining statements in the theorem we first remark that to each basis element $B_{\kappa\lambda}^\mu$ there corresponds at least one non-zero simple part \mathfrak{C}_{ij} of type (κ, λ) . For, by definition, $B_{\kappa\lambda}^\mu$ is a matrix of \mathfrak{A} having non-zero coefficients only in the upper left corners of simple parts of type (κ, λ) .

Let now $B_{\rho\sigma}^\tau$ be an element of the set \mathfrak{B} of Cartan basis elements which belong to a given block. If the simple parts of type (ρ, σ) are in part \mathfrak{B}_i of \mathfrak{A} , then by the above remark $B_{\rho\sigma}^\tau$ is an element of the invariant subalgebra \mathfrak{A}_i . If $B_{\mu\nu}^\gamma$ is also from the set \mathfrak{B} , that is, if $B_{\mu\nu}^\gamma$ belongs to the same block as $B_{\rho\sigma}^\tau$, then the simple parts of type (μ, ν) belong to the same block as simple parts of type (ρ, σ) . For a simple part of type (ρ, σ) may be connected by a chain of non-zero simple parts to a simple part of type (μ, ν) , the members of this chain corresponding to the members of the chain connecting $B_{\rho\sigma}^\tau$, $B_{\mu\nu}^\gamma$. Then the simple parts of type (μ, ν) are in the part \mathfrak{B}_i and so $B_{\mu\nu}^\gamma$ also is an element of \mathfrak{A}_i . It follows that the invariant subalgebra \mathfrak{A}^+ , generated by the elements of \mathfrak{B} , coincides with \mathfrak{A}_i . We now have at our disposal a correspondence between invariant subalgebras determined by blocks of Cartan basis elements on the one hand, and by blocks of simple parts on the other.

Now let

$$(13) \quad \mathfrak{A} = \mathfrak{A}_1 + \mathfrak{A}_2 + \cdots + \mathfrak{A}_q$$

be a direct decomposition of \mathfrak{A} into invariant subalgebras which may not be directly decomposed, and let \bar{A}_j be an element of \mathfrak{A}_j . If in the expression

$$\bar{A}_j = \sum h_{\alpha\beta}^\mu(\bar{A}_j) E_\kappa(a1) B_{\kappa\lambda}^\mu E_\lambda(1b)$$

for \bar{A}_j in terms of the Cartan basis elements, $h_{\alpha\beta}^\mu(\bar{A}_j) \neq 0$, then since

$$E_\kappa(1a) \bar{A}_j E_\lambda(b1) = h_{\alpha\beta}^\mu(\bar{A}_j) B_{\kappa\lambda}^\mu$$

is in \mathfrak{A}_j , so also is $B_{\kappa\lambda}^\mu$. In this way we can determine the direct summand of the decomposition (13) to which each element $B_{\kappa\lambda}^\mu$ belongs. In particular, we may determine the summands to which the elements $E_\kappa(11)$, $(\kappa = 1, 2, \dots, k)$ belong. If $B_{\kappa\lambda}^\mu$ belongs to \mathfrak{A}_j then $E_\kappa(11)$, $E_\lambda(11)$ must also belong to \mathfrak{A}_j , since

$$E_\kappa(11) B_{\kappa\lambda}^\mu = B_{\kappa\lambda}^\mu E_\lambda(11) = B_{\kappa\lambda}^\mu$$

and if, for instance, $E_\kappa(11)$ did not belong to the same summand as $B_{\kappa\lambda}^\mu$, the product $E_\kappa(11)B_{\kappa\lambda}^\mu$ would be zero.

Suppose now that the element $B_{\rho\sigma}^\gamma$ of the set \mathfrak{B} is in \mathfrak{A}_j but that there are elements of \mathfrak{B} which are not in \mathfrak{A}_j . Then we may choose an element $B_{\mu\nu}^\gamma$ of \mathfrak{B} which is not in \mathfrak{A}_j , but such that in the chain joining $B_{\rho\sigma}^\gamma$ to $B_{\mu\nu}^\gamma$, $B_{\mu\nu}^\gamma$ is the only member which does not belong to \mathfrak{A}_j . The member preceding $B_{\mu\nu}^\gamma$ in this chain has either the type index μ or ν ; let this member be $B_{\mu\lambda}^\delta$. Then by the above, $E_\mu(11)$ is in \mathfrak{A}_j , since $B_{\mu\lambda}^\delta$ is in \mathfrak{A}_j ; on the other hand $E_\mu(11)$ must be in the same summand as $B_{\mu\nu}^\gamma$, which gives a contradiction. Thus all elements of the set \mathfrak{B} are in \mathfrak{A}_j . Similarly, if any member of a second set \mathfrak{B}_1 of Cartan basis elements which all belong to one block is in \mathfrak{A}_j , then all members of the set \mathfrak{B}_1 are in \mathfrak{A}_j . But then the invariant subalgebras \mathfrak{A}^+ , \mathfrak{A}_1^+ which have as bases the sets \mathfrak{B} , \mathfrak{B}_1 , respectively, would be direct summands of \mathfrak{A} ; contrary to our assumption that \mathfrak{A}_j is directly indecomposable. We now have $\mathfrak{A}_i = \mathfrak{A}^+ = \mathfrak{A}$, which completes our proof.

We suppose now that each \mathfrak{B}_i has been split into its indecomposable parts¹³ and that these are in reduced form. Let $U_1^i, U_2^i, \dots, U_n^i$ be the indecomposable parts of \mathfrak{B}_i . We now classify the U^i by the following relation: we say that U_ρ^i, U_σ^i belong to the same block if there exists a chain¹⁴

$$(14) \quad U_\rho^i, U_\alpha^i, \dots, U_\beta^i, U_\sigma^i$$

such that any two neighboring parts of the chain (14) have at least one irreducible constituent in common. We will show that all indecomposable parts U_ρ^i ($\rho = 1, 2, \dots, n$) of \mathfrak{B}_i belong to one block. Let us take together all U_ρ^i that are connected with U_1^i by such a chain and suppose U_α^i does not belong to this set. Then the simple parts of U_α^i cannot be connected by a proper chain of non-zero simple parts of U_1^i , contrary to our provision that simple parts of \mathfrak{B}_i all belong to one block.

Since a matrix commuting with an indecomposable part U_ρ^i can have just one distinct characteristic root¹⁵ and since, as we have seen, all U_ρ^i of \mathfrak{B}_i have at least one irreducible constituent in common, then for an element Z of the center of \mathfrak{A} the matrix $B_i(Z)$ of \mathfrak{B}_i corresponding to Z has just one distinct characteristic root.

We have then proved

THEOREM 12: *The indecomposable constituent of a part \mathfrak{B}_i may be characterized as belonging to a block. It follows that for a center element Z , the corresponding matrix $B_i(Z)$ has only one distinct characteristic root.*

UNIVERSITY OF ALABAMA

¹³ For definitions of indecomposable parts, see B.N., or N.R.

¹⁴ R. Brauer and C. Nesbitt, *On the Modular Representations of Groups of Finite Order*, Univ. of Toronto Studies, Math. Series, No. 4 (1937), p. 14.

¹⁵ R. Brauer and I. Schur, *Zum Irreduzibilitätsbegriff in der Theorie der Gruppen linearer homogener Substitutionen*, Sitzber. der Preuss. Akad. der Wiss., Phys-Math. Klasse, Berlin, XIV (1930).

QUADRATIC FORMS PERMITTING COMPOSITION¹

By A. A. ALBERT

(Received November 8, 1941)

1. Introduction

An associative algebra, with a unity quantity, over a field \mathfrak{F} has an *involution*² (involutorial anti-automorphism) J such that $x + x'$ and xx' are in \mathfrak{F} if it is either \mathfrak{F} , an algebra of order two over \mathfrak{F} , a purely inseparable field the squares of whose elements are in \mathfrak{F} of characteristic two, or a generalized quaternion algebra over \mathfrak{F} . The quaternion algebras have order four and may be imbedded in a generalized Cayley-Dickson³ algebra of order eight over \mathfrak{F} . All the algebras so obtained are then *alternative*⁴ algebras with a unity quantity and an involution J such that $x + x'$ and xx' are in \mathfrak{F} . Then the *norm* xx' is a quadratic form in the coordinates of x which permits composition.

In 1898 A. Hurwitz⁵ showed that if \mathfrak{F} is the field \mathbb{C} of all complex numbers, forms equivalent to the quadratic norm forms described above are the only quadratic forms permitting composition. L. E. Dickson⁶ pointed out the connection of these forms with the corresponding algebras and thus completed the proof in the case $\mathfrak{F} = \mathbb{C}$ of the following

THEOREM. *A quadratic form over \mathfrak{F} permits composition if and only if it is equivalent in \mathfrak{F} to the norm form xx' of an alternative algebra over \mathfrak{F} with a unity quantity and an involution J such that $x + x'$ and xx' are in \mathfrak{F} . These latter **norm** forms are quadratic forms in 1, 2, 4 or 8 indeterminates except for the diagonal norm forms, in 2^i indeterminates, of purely inseparable fields of degree 2^i and exponent two over \mathfrak{F} of characteristic two.*

It is readily verified that Dickson's version of the Hurwitz proof is valid for any algebraically closed field of characteristic not two. However, there seems to be no treatment of the case \mathfrak{F} of characteristic two, a case presenting interesting new features. We shall derive the results for this case here and shall indeed

¹ Presented to the Society Nov. 21, 1941.

² An involution over \mathfrak{F} is a linear transformation J over \mathfrak{F} of an algebra \mathfrak{A} such that $(ab)^J = b^J a^J$ and J^2 is the identity transformation.

³ For the definition and elementary properties of such algebras see L. E. Dickson, *On quaternions and their generalizations and the history of the eight square theorem*, these Annals, vol. 20 (1919), pp. 155-71.

⁴ See M. Zorn, *Theorie der Alternativen Ringe*, Hamburg Abh. vol. 8 (1930), pp. 123-47 and his *Alternativkörper und Quadratische Systeme*, loc. cit. vol. 9 (1933), pp. 395-402 for a bibliography and the discussion of such algebras.

⁵ Göttingen Nachrichten, 1898, pp. 309-16.

⁶ See footnote 3.

provide an elegant unified study⁷ for the case where \mathfrak{F} is arbitrary. Our results will include a generalization⁸ of the Cayley-Dickson algebras to provide new algebras of order 2^e over \mathfrak{F} . Certain of these algebras \mathfrak{B} have connected quadratic norm forms and the property that if these norm forms are not null forms then every non-scalar quantity of \mathfrak{B} defines a quadratic subfield. However \mathfrak{B} need not then be a division algebra⁹ unless $e < 4$.

2. The general Hurwitz problem

Let $x_1, \dots, x_n, y_1, \dots, y_m$ be independent indeterminates over a field \mathfrak{F} , $\mathfrak{S} = \mathfrak{F}(x_1, \dots, x_n, y_1, \dots, y_m)$ be the corresponding rational function field. Designate the transpose of any matrix S with elements in \mathfrak{S} by S' and define the one rowed matrices

$$x = (x_1, \dots, x_n), \quad y = (y_1, \dots, y_m).$$

Then if A is any n -rowed square matrix with elements in \mathfrak{F} the matrix product $f(x) = xAx'$ is a quadratic form in x_1, \dots, x_n . Conversely every quadratic form $f(x)$ is expressible as such a product. However A is not unique and $xAx' = xA_0x'$ if and only if $A - A_0$ is an alternate (that is, skew-symmetric) matrix.

Suppose that B is an m -rowed square matrix so that $g(y) = yBy'$ is a second quadratic form. Then the general Hurwitz problem¹⁰ is that of determining under what conditions on $f(x)$ and $g(y)$ there exist quantities in \mathfrak{F} such that

$$(1) \quad f(x)g(y) = f(z),$$

where $z = (z_1, \dots, z_n)$ and

$$(2) \quad z_k = \sum_{i=1, \dots, n}^{j=1, \dots, m} x_i \gamma_{ik}^{(j)} y_j \quad (k = 1, \dots, n).$$

Write $G_j = (\gamma_{ik}^{(j)})$, so that G_j is an n -rowed square matrix with elements in \mathfrak{F} . Define

⁷ The argument required to make the extension of our theorem from the case where our field of reference is an algebraically closed field of characteristic not two to that where it is an arbitrary field seems not to be in the literature and will be a major contribution of our discussion.

⁸ The formulation by L. E. Dickson of the Cayley algebras is sufficiently general so that we are able to use it to define such algebras over a field of characteristic two. These algebras are new, however, and so are the other more general types (for example, of order 16) yielded by the Dickson process and given here.

⁹ We define algebras which are generalizations of the Cayley-Dickson algebras to algebras of order 2^t and with an involution J such that $x + x'$ and $xx' = x'x$ are in \mathfrak{F} . For proper choice of the defining parameters xx' is a quadratic form in 2^t squares. If \mathfrak{F} is the field of all real numbers each element of \mathfrak{B} defines a subalgebra equivalent to the complex field. But \mathfrak{B} is not a division algebra for $t > 3$. We shall show this in a later paper where the conditions that our algebras be division algebras will be discussed.

¹⁰ See his posthumous paper in the Math. Annalen, vol. 88 (1922), pp. 1-25.

$$(3) \quad G_y = G_1 y_1 + \cdots + G_m y_m,$$

so that if $e_i = (0, 0, \dots, 1, 0, \dots, 0)$ with 1 in the i^{th} column then

$$(4) \quad G_{e_i} = G_i.$$

We see that (2) is equivalent to the statement that z is the matrix product

$$(5) \quad z = x G_y.$$

Then (1) holds if and only if $x[g(y)A]x' = x G_y A (G_y)' x'$, that is, if and only if

$$(6) \quad N = g(y)A - G_y A (G_y)'$$

is an alternate matrix. Thus (2) holds if and only if

$$(7) \quad g(y)E = G_y E (G_y)', \quad E = A + A',$$

and the diagonal elements of N are all zero.

Let $\phi(\xi) = \xi C \xi'$ and $\psi(\zeta) = \zeta D \zeta'$ be equivalent to $f(x)$ so that there exist non-singular matrices P and Q such that $\xi C \xi' = \xi P A P' \xi'$, $\zeta D \zeta' = \zeta Q A Q' \zeta'$. Then $N_1 = C - P A P'$ and $N_2 = D - Q A Q'$ are alternate matrices. If (1) and (2) hold the matrix N of (6) is alternate and so is $P N P' = g(y) P A P' - \Gamma_y Q A Q' (\Gamma_y)'$, where $\Gamma_y = P G_y Q^{-1}$. Then the matrix

$$N_0 = g(y)C - \Gamma_y D \Gamma_y' = P N P' + g(y)N_1 - \Gamma_y N_2 (\Gamma_y)'$$

is also alternate. Let also $\rho(\eta)$ be equivalent to $g(y)$ so that there exists a non-singular matrix S such that $g(\eta S) = \rho(\eta)$. Then if $\Delta_\eta = \Gamma_y$, $y = \eta S$, the matrix

$$\rho(\eta)C - \Delta_\eta D (\Delta_\eta)' = N_0$$

is alternate and

$$(8) \quad \phi(\xi)\rho(\eta) = \psi(\zeta)$$

for $\zeta = (\zeta_1, \dots, \zeta_n)$ and the ζ_k bilinear forms with coefficients in \mathfrak{F} in the ξ_i and the η_j . Thus we have shown that (1) is possible if and only if (8) is possible for quadratic forms $\phi(x)$ and $\psi(x)$ equivalent to $f(x)$ and $\rho(\eta)$ equivalent to $g(y)$.

The principal purpose of our discussion is the derivation of a complete solution of the Hurwitz problem in the case $m = n$ and $g(x)$ equivalent to $f(x)$. Thus we shall stress the study of

$$(9) \quad f(x)f(y) = f(z),$$

for z as in (2) with $m = n$. If $f(x)$ has the property (9) we shall say that *the quadratic form $f(x)$ permits composition*. Then our principal object will be that the determination of the conditions on $f(x)$ that it permit composition.

3. The non-singular case

Let $f(x)$ be a quadratic form given by

$$(10) \quad f(x) = \sum_{i,k=1}^n x_i \alpha_{ik} x_k \quad (\alpha_{ik} \text{ in } \mathfrak{F}),$$

so that $f(x) = xAx'$ for A the n -rowed square matrix (α_{ik}) . If the characteristic of \mathfrak{F} is not two we may take A to be a non-singular diagonal matrix so that $\alpha_{ik} = 0$ for $i \neq k$, $\alpha_{ii} \neq 0$. However, when \mathfrak{F} has characteristic two the rank of $A + A'$ is an even integer $2r$, and $n \geq 2r$. Then¹¹

$$(11) \quad f(x) = \alpha_1 x_1^2 + \cdots + \alpha_n x_n^2 \quad (\alpha_i \neq \text{in } \mathfrak{F})$$

if $r = 0$, and otherwise

$$(12) \quad f(x) = \alpha_1 x_1^2 + \cdots + \alpha_n x_n^2 + (x_1 x_{r+1} + \cdots + x_r x_{2r}).$$

Here we assume that if $n > 2r$ the quantities $\alpha_{2r+1}, \dots, \alpha_n$ are all not zero. Moreover if $0 < i < r$ and both α_i and α_{i+r} are zero the transformation $x_i = \xi_i$, $x_{i+r} = \xi_i + \xi_{i+r}$ gives $x_i x_{i+r} = \xi_i^2 + \xi_i \xi_{i+r}$. If $\alpha_i = 0$, $\alpha_{i+r} \neq 0$ we may interchange the corresponding indeterminates. Thus we may assume in all cases that if $r > 0$ the quantities $\alpha_1, \dots, \alpha_r$ are all not zero.

If \mathfrak{F} has characteristic not two and A is a non-singular diagonal matrix as above then $E = A + A' = 2A$ is non-singular. However when \mathfrak{F} has characteristic two the matrix E is non-singular if and only if $n = 2r$,

$$(13) \quad A = \begin{pmatrix} D_1 & I_r \\ 0 & D_2 \end{pmatrix}, \quad E = A + A' = \begin{pmatrix} 0 & I_r \\ I_r & 0 \end{pmatrix},$$

where D_1 and D_2 are diagonal matrices and D_1 is non-singular. We shall show later that a quadratic form $f(x)$ permits composition only if it is either a diagonal form (11) or is equivalent to a form xAx' with $A + A'$ non-singular. Let us then study (1) temporarily only in this *non-singular case*, and treat the remaining cases later.

If we replace $y = (y_1, \dots, y_m)$ in $g(y)$ by $\eta = (\eta_1, \dots, \eta_m)$ for the η_i in \mathfrak{F} we obtain a quantity $\gamma = g(\eta)$ in \mathfrak{F} which we say is represented by $g(y)$. Then we have

LEMMA 1. *Let $\gamma \neq 0$ in \mathfrak{F} be any quantity represented by $g(y)$ in the non-singular case of (1). Then $f(x)$ is equivalent to $\gamma^{-1}f(x)$.*

For by (1) we have $f(x)\gamma = f(u)$ where $u = xP$ and $P = G_\eta$. Thus $f(x) = \gamma^{-1}f(u)$ is equivalent to $\gamma^{-1}f(x)$ if we can show that the matrix P is non-singular. But $x\gamma Ax' = xPAP'x'$ if and only if $\gamma A - PAP'$ is alternate. Hence $\gamma E = PEP'$ and P must be non-singular if $\gamma \neq 0$ and E is non-singular.

¹¹ For these results see Chapter II of my *Symmetric and alternate matrices in an arbitrary field*, Transactions of the A. M. S., vol. 43 (1938), pp. 386-436.

Write

$$(14) \quad g(y) = \sum_{i,j=1}^m y_i \beta_{ij} y_j,$$

where we are taking $\beta_{11} \neq 0$. Then (1) implies that $\phi(x)g(y) = \phi(z)$ where $\phi(x) = \alpha_{11}^{-1}f(x)$. Also $g(y) = \beta_{11}\psi(y)$ and $\beta_{11} = g(\eta)$, $\eta = (1, 0, \dots, 0)$. It follows that $\phi(x)$ is equivalent to $\beta_{11}^{-1}\phi(x)$ and that then $[\beta_{11}^{-1}\phi(x)] \cdot [\beta_{11}\psi(y)] = \phi(z)$, $\phi(x)\psi(y) = \phi(z)$. Thus we have shown that in the non-singular case there is no loss of generality if we assume that $\alpha_{11} = \beta_{11} = 1$.

Note that in the study of quadratic forms $f(x)$ permitting composition Lemma 1 implies that $f(x)$ is equivalent to $\alpha_{11}^{-1}f(x)$ and thus that we need only study forms $f(x)$ with $\alpha_{11} = 1$ in the non-singular case.

We now write (7) in the equivalent form

$$(15) \quad G_y(G_y)^J = g(y)I,$$

where I is the n -rowed identity matrix and

$$(16) \quad (G_y)^J = EG_yE^{-1} = G_1^J y_1 + \dots + G_m^J y_m.$$

It is well known that the correspondence $G \rightarrow G^J = EG'E^{-1}$ is an involution of the algebra of all n -rowed square matrices with elements in \mathfrak{S} , and we shall use the consequent properties of J .

The equation (7) is an identity in the indeterminates y_1, \dots, y_m and is equivalent, in view of (14), to

$$(17) \quad G_i G_i^J = \beta_{ii} I, G_i G_j^J + G_j G_i^J = \delta_{ij} I \quad (i \neq j; i, j = 1, \dots, m),$$

where

$$(18) \quad \delta_{ij} = \delta_{ji} = \beta_{ij} + \beta_{ji}.$$

In particular $G_1 G_1^J = I$ since we have taken $\beta_{11} = 1$. Then also $G_1^J G_1 = I$. We now define

$$(19) \quad G_y^{(0)} = G_1^J G_y = G_1^{(0)} y_1 + \dots + G_m^{(0)} y_m$$

and clearly have

$$(20) \quad G_1^{(0)} = I, G_y^{(0)} (G_y^{(0)})^J = G_1^J (G_y G_y^J) G_1 = g(y) G_1^J G_1 = g(y) I,$$

so that $G_y^{(0)}$ is a solution of (7) if G_y is. But also $G_1^J = G_y^{-1}$ and $g(e_1) = 1$, (6) states that $A - G_1 A G_1^J$ is alternate and so is $G_1^J A (G_1^J)' - A$. Then by (6)

$$N_0 = [A - G_1^J A (G_1^J)'] g(y) + G_1^J [g(y) A - G_y A (G_y)'] (G_1^J)' = A g(y) - G_y^{(0)} A (G_y^{(0)})'$$

is an alternate matrix. It follows that G_y defines a solution (5) of (2) so that (1) holds if and only if $G_y^{(0)}$ does, and we may thus assume that

$$(21) \quad G_1 = G_1^J = I.$$

Equations (17) for $i = 1$ now give

$$(22) \quad G_j^J = \delta_{1j}I - G_j \quad (j = 2, \dots, m),$$

while $G_1^J = 2I - G_1$. But then

$$(23) \quad (G_y)^J = t(G_y) - G_y,$$

where the trace function, $t(G_y)$, is the scalar matrix

$$(24) \quad G_y + (G_y)^J = \left(2y_1 + \sum_{i=2}^m \delta_{1i}y_i\right)I,$$

and is linear in the coordinates y_i of y . Thus the mapping

$$\eta \rightarrow G_\eta$$

is a linear mapping of the linear space \mathfrak{L} of one by m vectors η on the linear space of matrices G_η . Moreover if we define $e = (1, 0, \dots, 0)$,

$$(25) \quad \eta^J = \left(2\eta_1 + \sum_{i=2}^m \delta_{1i}\eta_i\right)e - \eta$$

we see that J is a linear transformation on \mathfrak{L} such that

$$(26) \quad G_{\eta^J} = (G_\eta)^J.$$

But $[(G_\eta)^J]^J = G_\eta$ and hence

$$(27) \quad (\eta^J)^J = \eta.$$

We shall now specialize our study of (1) to the case $m = n$ and $g(y) = f(y)$.

4. Composition in the non-singular case

Let $f(x)$ be a quadratic form permitting composition. Then

$$(28) \quad f(x)f(y) = f(xG_y),$$

where the elements of the n -rowed square matrix G_y are linear forms in y_1, \dots, y_n and we have seen that if $e = (1, 0, \dots, 0)$ we may take

$$(29) \quad G_e = I, \quad f(e) = 1, \quad G_y G_y^J = f(y)I$$

such that (23) and (24) hold. Then the ordinary matrix product $z = xG_y$ gives (2) and these equations may also be written as the first equation in

$$(30) \quad z = yH_x = xG_y,$$

where we define $\rho_{jk}^{(i)} = \gamma_{ik}^{(j)}$ and have

$$(31) \quad H_i = (\rho_{jk}^{(i)}), \quad H_x = H_1x_1 + \dots + H_nx_n.$$

But $f(y)f(x) = f(yH_x)$ and our derivation of the properties of G_y implies that we may interchange the roles of $f(x)$ and $f(y)$ to obtain

$$(32) \quad H_x H_x^J = f(x)I,$$

as a consequence of (15). It follows from (29) that

$$(33) \quad H_e H_e' = I,$$

and thus that H_e is non-singular. Define

$$(34) \quad u = u(y) = yH_e, \quad R_u = G_y,$$

so that $u = (u_1, \dots, u_n)$, $y = uH_e^{-1}$ and y_1, \dots, y_n are linear forms in u_1, \dots, u_n . Then by (30) and (28) with $x = e$ we have

$$(35) \quad u = eG_y = eR_u, \quad f(u) = f(eG_y) = f(e)f(y) = f(y).$$

Moreover $u(e) = eG_e = e$ since $G_e = I$, $R_{u(e)} = G_e = I$. But then $f(x)f(y) = f(xG_y) = f(xR_u) = f(x)f(u)$,

$$(36) \quad f(x)f(y) = f(xR_y).$$

We have thus proved that R_y is a solution G_y of (2) such that

$$(37) \quad R_e = I, \quad eR_y = y, \quad R_y R_y' = f(y)I, \quad N = f(y)A - R_y A R_y'$$

is an alternate matrix.

Let now \mathfrak{L} be the linear space of all $\xi = (\xi_1, \dots, \xi_n)$ for ξ_i in \mathfrak{F} . Then the correspondence

$$(38) \quad \xi \rightarrow R_\xi$$

is evidently a linear mapping of \mathfrak{L} on the linear space of all the matrices R_ξ . Define an operation

$$(39) \quad \xi \cdot \eta = \xi R_\eta$$

on \mathfrak{L} to \mathfrak{L} and call the result $\xi \cdot \eta$ the *product* of ξ and η . Then \mathfrak{L} becomes an algebra \mathfrak{A} of order n over \mathfrak{F} . Indeed all algebras may be defined in this way.

We have $e \cdot \eta = eR_\eta = \eta$ by (37), and $\eta \cdot e = \eta R_e = \eta I = \eta$. Hence \mathfrak{A} has e as its unity quantity. We shall give further properties of \mathfrak{A} in the next section. We now relate this result to a certain class of algebras.

An algebra \mathfrak{A} over \mathfrak{F} is called *alternative* if

$$(40) \quad x(xy) = (xx)y, \quad (yx)x = y(xx)$$

for every x and y of \mathfrak{A} . Then $(x + y)[(x + y)x] = (x + y)(xx + yx) = x(xx) + x(yx) + y(xx) + y(yx) = [(x + y)(x + y)]x = (xx + xy + yx + yy)x = (xx)x + (xy)x + (yx)x + (yy)x$. By (40) we have $x(yx) = (xy)x$, a postulate which is sometimes given as one of the alternative postulates but which is actually a consequence of (40). We now prove

LEMMA 2. *Let \mathfrak{A} be an algebra with a unity quantity and an involution J such that*

$$(41) \quad t_x = x + x', \quad xx'$$

are in \mathfrak{F} . Then $x'x = xx'$, and \mathfrak{A} is alternative if and only if

$$(42) \quad (xx')y = x(x'y) = x'(xy) = (yx)x' = (yx')x.$$

For $xx' = x(t_x - x) = t_x x - xx = (t_x - x)x = x'x$ since t_x is a scalar. Also

$$x(x'y) = x[(t_x - x)y] = t_x(xy) - x(xy),$$

and

$$(xx')y = [x(t_x - x)]y = t_x(xy) - (xx)y.$$

Thus $x(x'y) = (xx')y$ if and only if $x(xy) = (xx)y$. Similarly $(yx)x' = y(xx')$ if and only if $(yx)x = y(xx)$. However $(x')' = x$ and if \mathfrak{A} is alternative we have $x'(xy) = (x'x)y = (xx')y$, $(yx')x = y(x'x) = (xx')y$.

We next prove

LEMMA 3. *Let \mathfrak{A} be alternative and as in Lemma 1. Then the quadratic form xx' in the coordinates x_1, \dots, x_n of the quantities $x = (x_1, \dots, x_n)$ of \mathfrak{A} permits composition.*

For it is clear that if \mathfrak{S} is any scalar extension of \mathfrak{F} the algebra $\mathfrak{A}_{\mathfrak{S}}$ has the properties we have assumed for \mathfrak{A} . We let $x_1, \dots, x_n, y_1, \dots, y_n$ be independent indeterminates over \mathfrak{F} and $\mathfrak{S} = \mathfrak{F}(x_1, \dots, x_n, y_1, \dots, y_n)$. Then $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)$ are non-zero quantities of $\mathfrak{A}_{\mathfrak{S}}$ and so are x', y' . Also $xx' = x'x = f(x)$ is in \mathfrak{S} . Then so are $f(y) = yy' = y'y$, $f(xy) = (xy)(xy)'$. But $[f(xy)]x' = x'[(xy)(xy)'] = [x'(xy)](y'x') = f(x)[y(y'x')] = f(x)f(y)x', f(x)f(y) = f(xy)$ as desired.

We now prove the partial converse,

LEMMA 4. *Let $f(x) = xAx'$ such that $A + A'$ is non-singular. Then $f(x)$ permits composition if and only if $f(x)$ is the norm form xx' of an alternative algebra \mathfrak{A} with a unity quantity and an involution J such that $x + x'$ and xx' are¹² in \mathfrak{F} .*

For $f(x) = \sum x_i \alpha_{ij} x_j$ with $\alpha_{11} = 1$ and all the α_{ij} in \mathfrak{F} and we define an algebra \mathfrak{A} over \mathfrak{F} by (39) and have (37). Then if $\xi = (\xi_1, \dots, \xi_n)$ with the ξ_i in \mathfrak{F} we have seen that $R_{\xi} + (R_{\xi})' = t_{\xi}I$ where $t = t_{\xi}$ is in \mathfrak{F} and is determined by $f(x)$, that is, actually

$$(43) \quad t_{\xi} = 2\xi_1 + \sum_{j=2}^n (\alpha_{1j} + \alpha_{j1})\xi_j.$$

We now write

$$(44) \quad \xi' = t_{\xi}e - \xi,$$

and have defined a correspondence J which is evidently a linear transformation on \mathfrak{A} . Now if $t = t_{\xi}$ we have

$$(45) \quad R_{\xi'} = R_{t_{\xi}e - \xi} = tI - R_{\xi} = (R_{\xi})'$$

so that $t_{\xi'} = t_{\xi}$, $(\xi')' = \xi$ for every ξ of \mathfrak{A} . Since $\xi\eta$ is not defined as a matrix product we may drop the dot in (39). We define $\mathfrak{S} = \mathfrak{F}(x_1, \dots, x_n,$

¹² In fact we show that each solution G_y of the consequent equation $z = xG_y$ defines an algebra. Moreover those of these algebras which have unity quantities have the property of our lemma.

y_1, \dots, y_n) as above and in $\mathfrak{A}_{\mathfrak{F}}$ have $x = (x_1, \dots, x_n)$, $y = (y_1, \dots, y_n)$, $f(x)e = xx^J$ is the norm form of \mathfrak{A} as desired. By the proof of Lemma 1 $xx^J = x^Jx$. Finally

$$(46) \quad (yx)x^J = (yR_x)(R_x^J) = y(R_xR_x^J) = yf(x),$$

so that we have part of (42).

We note now that we may write $\xi\eta = \xi R_\eta = \eta L_\xi$ as in (30). But then $L_\xi + (L_\xi)^J = t_\xi I$ since t_ξ depends only on ξ_1, \dots, ξ_n and the coefficients of $f(x)$, not on the choice of the solution G_ν of (2). Thus $(L_\xi)^J = L_{\xi^J}$. It follows that $x^J(xy) = (yL_x)L_{x^J} = yL_x(L_x)^J = yf(x)$. We now use $f(x)f(y) = f(xy)$ and form $x^Jf(xy) = x^J[(xy)(xy)^J] = x^J(xy)(xy)^J = f(x)y(xy)^J = f(x)f(y)x^J$. Thus $y(xy)^J = (yy^J)x^J = y(y^Jx^J)$. Multiplying by y^J we have $y^Jy(xy)^J = (y^Jy)(y^Jx^J)$, $(xy)^J = y^Jx^J$. But then $(\xi\eta)^J = \eta^J\xi^J$ for every ξ and η of \mathfrak{A} , J defines an involution of \mathfrak{A} such that $\xi + \xi^J$ and $\xi\xi^J$ are in \mathfrak{F} , \mathfrak{A} is alternative as desired. This completes our proof.

5. Algebras and their norm forms

The proof of our principal theorem requires that we exhibit algebras \mathfrak{A} whose norm forms xx^J are quadratic forms xAx' with $A + A'$ non-singular. We shall show later that such algebras necessarily have orders 1, 2, 4, 8. Let us then exhibit algebras of the kind desired and of these orders.

The algebra of order one is \mathfrak{F} itself. Any algebra $\mathfrak{A} = (1, u)$ with $u^2 = \beta u + \gamma$ and β and γ in \mathfrak{F} has the desired properties and is trivially shown to be an associative algebra. Then if $\beta = 0$ and \mathfrak{F} has characteristic two we have $u^2 = \gamma$, $(u^J)^2 = \gamma$, $u^J = u$ for any involution J . But then $xx^J = xx = x_1^2 + x_2^2\gamma = xAx'$ with $A + A' = 0$, contrary to hypothesis. If $\beta = 0$ and \mathfrak{F} does not have characteristic two we have $(u + 1)^2 = u^2 + 2u + 1 = 2(u + 1) + \gamma - 1$ and may take $\beta \neq 0$. Define $k = \beta^{-1}u$ and have $k^2 = k + \alpha$ for α in \mathfrak{F} . Thus the algebra

$$(47) \quad \mathfrak{A} = (1, k), \quad k^2 = k + \alpha,$$

is the only algebra of order two which we need to consider. It is associative, has the automorphism T defined by

$$(48) \quad x = x_1 + x_2k, \quad x^T = x_1 + x_2(1 - k),$$

and T is an involution of \mathfrak{A} . It follows that

$$(49) \quad x + x^T = 2x_1 + x_2, \quad xx^T = x_1^2 + x_1x_2 - x_2^2.$$

The *quaternion* algebras may be defined as algebras \mathfrak{A} of matrices

$$(50) \quad x = \begin{pmatrix} a & b \\ b^T\beta & a^T \end{pmatrix},$$

with $\beta \neq 0$ in \mathfrak{F} , $a = x_1 + x_2k$, $b = x_3 + x_4k$, and k and T defined for the algebra

given by (47), (48). Then \mathfrak{A} has a basis $1, u, v, uv$ over \mathfrak{F} , where the unity quantity 1 of \mathfrak{A} is really the two rowed identity matrix, and

$$(51) \quad u = \begin{pmatrix} k & 0 \\ 0 & 1 - k \end{pmatrix}, \quad v = \begin{pmatrix} 0 & 1 \\ \beta & 0 \end{pmatrix}.$$

Then $vu = (1 - u)v$, $u^2 = u + 1$, $v^2 = \beta \neq 0$. Conversely every algebra with such a multiplication table is a simple associative algebra of degree two over its center¹³ \mathfrak{F} , and is representable as the set of matrices (50).

Every involution¹⁴ of a two-rowed total matrix algebra is a correspondence

$$(52) \quad x \rightarrow x^J = ex'e^{-1},$$

for a non-singular symmetric or alternate matrix e . We take

$$(53) \quad e = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix},$$

and see that

$$(54) \quad x^J = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} a & b^T\beta \\ b & a^T \end{pmatrix} \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} a^T & -b \\ -b^T\beta & a \end{pmatrix}.$$

Then x^J is in \mathfrak{A} and (54) defines an involution over \mathfrak{F} of \mathfrak{A} . Moreover if $a = x_1 + x_2k$ we have

$$(55) \quad x + x^J = a + a^T = 2x_1 + x_2$$

in \mathfrak{F} ,

$$(56) \quad f(x) = \det \begin{pmatrix} a & b \\ b^T\beta & a^T \end{pmatrix} = \det \begin{pmatrix} a^T & -b \\ -b^T\beta & a \end{pmatrix} = aa^T - \beta bb^T = x^J x.$$

Thus $f(x) = x_1^2 + x_1x_2 - \alpha x_2^2 - (x_3^2 + x_3x_4 - x_4^2\alpha)\beta$ is the determinant of x and is a multiplicative function, that is, $f(x)f(y) = f(z)$ where $z = xy$.

Forms in x_1, \dots, x_8 have been connected¹⁵ with certain non-associative alternative algebras. We shall construct these algebras over fields of arbitrary characteristic by the use of L. E. Dickson's formulation of the non-associative algebra of Cayley. Let \mathfrak{A} be an algebra of order n over \mathfrak{F} and let \mathfrak{A} have a unity quantity 1 which we shall identify with the unity quantity of \mathfrak{F} . Assume that \mathfrak{A} has an involution J over \mathfrak{F} such that $x + x^J$ and $xx^J = x^Jx$ are in \mathfrak{F} . Then every x of \mathfrak{A} is a root of an equation

$$\phi(\lambda) = (\lambda - x)(\lambda - x^J) = \lambda^2 - (x + x^J)\lambda + xx^J = 0$$

with coefficients $t(x) = x + x^J$ and $f(x) = xx^J$ in F . Hence \mathfrak{A} has degree two over \mathfrak{F} and $\phi(\lambda)$ is the minimum function of every x of \mathfrak{A} which is not in \mathfrak{F} .

¹³ We use *center* instead of *centrum*.

¹⁴ See Chapter V of my *Modern Higher Algebra*.

¹⁵ See the paper referred to in footnote 3.

Every x in \mathfrak{A} defines a subalgebra $\mathfrak{F}[x]$ of all polynomials in x and the unity quantity of \mathfrak{A} with coefficients in \mathfrak{F} . Clearly if x is not in \mathfrak{F} then $\mathfrak{F}[x] = (1, x)$ over \mathfrak{F} and $\mathfrak{F}[x]$ is an associative subalgebra of \mathfrak{A} . Since $x^J = t(x) - x$ we see that the involution J of \mathfrak{A} induces an automorphism in every subalgebra $\mathfrak{F}[x]$ of \mathfrak{A} which is the identity if and only if either x is in \mathfrak{F} or $t(x) = 0$ and \mathfrak{F} has characteristic two. Observe now that the function $f(x) = xx^J$ is a quadratic form in the coordinates of x with respect to any basis of \mathfrak{A} and that every $x \neq 0$ of \mathfrak{A} has an inverse in \mathfrak{A} and actually in $\mathfrak{F}[x]$ if and only if $f(x)$ is not a null form.

We shall now construct an algebra \mathfrak{B} of order $2n$ over \mathfrak{F} with the properties described above by the use of any such algebra \mathfrak{A} of order n over \mathfrak{F} . The algebra \mathfrak{B} will have quantities which are pairs of quantities of \mathfrak{A} and it is customary to indicate this by the statement that these quantities are uniquely expressible in the form

$$(57) \quad x = g_1 + g_2 w$$

for g_1 and g_2 in \mathfrak{A} . Let $\gamma \neq 0$ be in \mathfrak{F} , J be the given involution of \mathfrak{A} and define multiplication¹⁶ in \mathfrak{B} by

$$(58) \quad xy = (g_1 + g_2 w)(h_1 + h_2 w) = (g_1 h_1 + \gamma h_2^J g_2) + (h_2 g_1 + g_2 h_1^J) w.$$

Extend the definition of J so that J becomes a linear transformation of \mathfrak{B} by means of

$$(59) \quad x^J = g_1^J - g_2 w.$$

It is clear that J is now a one-to-one linear transformation of \mathfrak{B} such that J^2 is the identity transformation.

To form the product xx^J we replace h_2 by $-g_2$, h_1^J by g_1 in (58) and have $h_2 g_1 + g_2 h_1^J = -g_2 g_1 + g_2 g_1 = 0$. Then

$$(60) \quad xx^J = g_1 g_1^J - \gamma g_2 g_2^J = x^J x$$

since the interchange of g_1 with g_1^J , g_2 with $-g_2$ does not alter $g_1 g_1^J - \gamma g_2 g_2^J$. Also

$$(61) \quad x + x^J = g_1 + g_1^J$$

is in \mathfrak{F} , each x of \mathfrak{B} is a root of $\lambda^2 - (x + x^J)\lambda + xx^J = 0$, and \mathfrak{B} has degree two over \mathfrak{F} . Finally

$$(62) \quad (xy)^J = (h_1^J g_1^J + \gamma g_2^J h_2) - (h_2 g_1 + g_2 h_1^J) w,$$

and

$$(63) \quad \begin{aligned} y^J x^J &= (h_1^J - h_2 w)(g_1^J - g_2 w) \\ &= (h_1^J g_1^J + \gamma g_2^J h_2) - (g_2 h_1^J + h_2 g_1) w, \end{aligned}$$

¹⁶ This is clearly a generalization of the L. E. Dickson formulation of the definition of the Cayley algebra of order eight.

so that $(xy)^J = y^J x^J$ and J is an involution of \mathfrak{B} . We observe that (58) may be given more simply by the use of the distributive law and the relations

$$(64) \quad \begin{aligned} g(hw) &= (hg)w, & (gw)h &= (gh^J)w, \\ (gw)(hw) &= \gamma h^J g, \end{aligned}$$

for every g and h of \mathfrak{A} . Clearly if \mathfrak{A} is not a commutative algebra the algebra \mathfrak{B} is not associative.

To prove that \mathfrak{B} is alternative we use (57) to compute

$$(65) \quad \begin{aligned} x(xy) &= [g_1(g_1 h_1 + \gamma h_2^J g_2) + \gamma(g_1^J h_2^J + h_1 g_2^J)g_2] \\ &\quad + [(h_2 g_1 + g_2 h_1^J)g_1 + g_2(h_1^J g_1^J + \gamma g_2^J h_2)]w, \end{aligned}$$

as well as $x^2 = g_1^2 + \gamma g_2^J g_2 + [g_2(g_1 + g_1^J)]w$ and

$$(66) \quad \begin{aligned} x^2 y &= \{(g_1^2 + \gamma g_2^J g_2)h_1 + \gamma h_2^J [g_2(g_1 + g_1^J)]\} \\ &\quad + \{h_2(g_1^2 + \gamma g_2^J g_2) + [g_2(g_1 + g_1^J)]h_1^J\}w. \end{aligned}$$

If \mathfrak{A} is an alternative algebra we have $g_1(g_1 h_1) + \gamma(h_1 g_2^J)g_2 = g_1^2 h_1 + \gamma(g_2^J g_2)h_1$. However $g_1 + g_1^J$ is in \mathfrak{F} and $h_2^J [g_2(g_1 + g_1^J)] = (g_1 + g_1^J)(h_2^J g_2) = g_1(h_2^J g_2) + (g_1^J h_2^J)g_2$ if and only if $g_1^J(h_2^J g_2) = (g_1^J h_2^J)g_2$, that is, \mathfrak{A} is an associative algebra. Conversely if \mathfrak{A} is associative we have also $(h_2 g_1)g_1 + \gamma g_2(g_2^J h_2) = h_2(g_1^2 + \gamma g_2^J g_2)$, $(g_2 h_1^J)g_1 + g_2(h_1^J g_1^J) = g_2(g_1 + g_1^J)h_1^J$, and $x(xy) = x^2 y$ for every x and y of \mathfrak{B} .

It follows that $x^J(x^J y^J) = (x^J x^J)y^J$ and, by operating with J , that $(yx)x = y(xx)$ for every x and y of \mathfrak{B} . Hence \mathfrak{B} is alternative. This completes our proof of the result that \mathfrak{B} is an alternative algebra if and only if \mathfrak{A} is associative.

We now see that if \mathfrak{A} is the associative algebra of order four defined by the matrices of (50) the algebra \mathfrak{B} is an alternative algebra of order eight over \mathfrak{F} . Its norm form $f(x) = xx^J = x^J x$ is given explicitly by

$$(67) \quad \begin{aligned} f(x) &= x_1^2 + x_1 x_2 - x_2^2 \alpha - (x_3^2 + x_3 x_4 - x_4^2 \alpha) \beta \\ &\quad - (x_5^2 + x_5 x_6 - x_6^2 \alpha) \gamma + (x_7^2 + x_7 x_8 - x_8^2 \alpha) \beta \gamma. \end{aligned}$$

Note that $f(x)$ becomes the norm form of the algebra \mathfrak{A} of (50) by putting $x_5 = x_6 = x_7 = x_8 = 0$, and that of (47) by taking also $x_3 = x_4 = 0$. The values $\alpha = \beta = \gamma = -1$ then give us the forms $x_1^2 + x_2^2 + x_1 x_2$, $x_1^2 + \dots + x_4^2 + x_1 x_2 + x_3 x_4$, $x_1^2 + \dots + x_8^2 + x_1 x_2 + x_3 x_4 + x_5 x_6 + x_7 x_8$.

Observe that the algebra \mathfrak{B} of order eight defined above is a division algebra if and only if its associative subalgebra \mathfrak{A} is a division algebra and $f(x)$ is not a null form. For if $xy = 0$ then $x^J(xy) = f(x)y = 0$ and if $x \neq 0$ then $f(x) \neq 0$, $y = 0$. The condition that $xx^J = 0$ if and only if $x = 0$ is satisfied when γ is chosen so that γ is not the norm gg^J of any g of \mathfrak{A} .

It is already known that in \mathfrak{F} of characteristic not two a quadratic form $f(x) = xAx'$ with $A = A'$ a non-singular n -rowed matrix permits composition only when $n = 1, 2, 4, 8$. Thus it remains to show the corresponding result for

$A + A'$ non-singular and \mathfrak{F} of characteristic two, as well as to study other possible cases of composition for this characteristic. Let us thus assume henceforth that the characteristic of \mathfrak{F} is two.

6. Determination of possible orders in the non-singular case

Let \mathfrak{A} be any simple algebra of degree n over its center \mathfrak{F} and let I be the unity quantity of \mathfrak{A} . Assume the existence of quantities U and V in \mathfrak{A} such that

$$(68) \quad U^2 = \beta I, \quad V^2 = \gamma I, \quad W = UV = I - VU,$$

for $\beta \neq 0$ and γ in \mathfrak{F} . Then $W^2 = U(I - UV)V = W - \beta\gamma I$, $UW + WU = \beta V + U(I - UV) = U$. It follows that

$$(69) \quad W^2 = W - \beta\gamma I, \quad UW = (I - W)U.$$

But then \mathfrak{A} contains a generalized quaternion algebra $\mathfrak{B} = (I, W, U, WU) = (I, U, V, UV)$. This algebra of order four over \mathfrak{F} is a simple algebra of degree two over its center \mathfrak{F} . But then it is well known¹⁷ that \mathfrak{A} is the direct product

$$(70) \quad \mathfrak{A} = \mathfrak{B} \times \mathfrak{A}_1, \quad \mathfrak{A}_1 = \mathfrak{A}^{\mathfrak{B}},$$

where we write $\mathfrak{A}^{\mathfrak{B}}$ for the set of all quantities of \mathfrak{A} commutative with both U and V . Moreover $n = 2n_1$, \mathfrak{A}_1 is a simple algebra of degree n_1 over its center \mathfrak{F} . Assume now that \mathfrak{A} contains also U_2, V_2 such that $U_2^2 = \beta_2 I$, $V_2^2 = \gamma_2 I$, $U_2 V_2 + V_2 U_2 = I$, where $\beta_2 \neq 0$ and γ_2 are in \mathfrak{F} . Then if $U U_2 - U_2 U = U V_2 - V_2 U = V U_2 - U_2 V = V V_2 - V_2 V = 0$, the quantities U_2 and V_2 are in $\mathfrak{A}^{\mathfrak{B}}$. But then \mathfrak{A}_1 contains $\mathfrak{B}_2 = (I, U_2, V_2, U_2 V_2)$, and $\mathfrak{A}_1 = \mathfrak{B}_2 \times \mathfrak{A}_2$ where $n_1 = 2n_2$, $\mathfrak{A}_2 = \mathfrak{A}^{\mathfrak{B}_2}$ is simple of degree n_2 over its center \mathfrak{F} . We now let $s \geq 1$ and suppose that \mathfrak{A} contains $s - 1$ pairs of quantities U_j, V_j such that

$$(71) \quad U_j^2 = \beta_j I, \quad V_j^2 = \gamma_j I, \quad U_j V_j + V_j U_j = I \quad (j = 1, \dots, s - 1),$$

for $\beta_j \neq 0$ and γ_j in \mathfrak{F} . Assume also that

$$(72) \quad U_j U_k - U_k U_j = U_j V_k - V_k U_j = V_j V_k - V_k V_j = 0$$

$$(j \neq k; j, k = 1, \dots, s - 1).$$

Then $\mathfrak{A} = \mathfrak{B}_1 \times \dots \times \mathfrak{B}_{s-1} \times \mathfrak{A}_s$, and we have proved that 2^{s-1} divides n .

Let us now apply these results to the study of (1) in the case where $f(x) = xAx'$, $E = A + A'$ is non-singular, and we have (12) for $n = 2r$, $\alpha_1 = 1$. We assume that \mathfrak{F} has characteristic two and that

$$(73) \quad g(y) = y_1^2 + \beta_2 y_2^2 + \dots + \beta_m y_m^2 + (y_1 y_{s+1} + \dots + y_s y_{2s})$$

where $m \geq 2s$, the β_i are in \mathfrak{F} , and β_2, \dots, β_s are all not zero. By (17) there exist matrices $G_1 = I, G_2, \dots, G_m$ such that

$$(74) \quad G_i G_i' = \beta_i, \quad G_i G_k' + G_k G_i' = \delta_{ik} I \quad (i \neq k; i, k = 1, \dots, m),$$

¹⁷ See my *Structure of Algebras*, pp. 52-55.

where now $\delta_{ik} = 0$ unless $0 < i \leq 2s$, $0 < k \leq 2s$ and $i - k$ is s or $-s$. In the remaining cases $\delta_{ik} = 1$. Define $U_{j-1} = G_j$, $V_{j-1} = G_{j+s}$ for $j = 2, \dots, s$. Then by (74) for $i = 1$ we have $G'_k + G_k = 0$ for $k = 2, \dots, s$ and thus $G'_k = G_k$. Hence $U_j^2 = \beta_{j+1}$, $V_j^2 = \beta_{j+1+s}$ for $j = 1, \dots, s-1$ where $\beta_{j+1} \neq 0$. Also by (74) for $i = j+1$ and $k = i+s$ we have $U_j V_j + V_j U_j = I$ and thus (71). By these same equations for $k - i \neq s$ and $i = j+1$ or $j+1+s$ respectively we obtain (72). We let \mathfrak{A} be the algebra of all n -rowed square matrices with elements in \mathfrak{F} and it follows as above that 2^{s-1} divides n .

Note that we have not utilized all consequences of (74) and that the conditions omitted as well as other results would have to be used in order to complete the study of (1). However we are using this study of (1) only as a tool in our discussion of quadratic forms permitting composition and so shall leave the more general considerations for later study.

Let us then take $m = 2s = n = 2r$. The results just derived imply that 2^{r-1} divides $2r$. This is not the case for $r = 3$, $r > 4$, it is the case for $r = 1, 2, 4$, and thus $n = 2, 4, 8$. We have shown that in the non-singular case over \mathfrak{F} of characteristic two a form $f(x) = f(x_1, \dots, x_n)$ permits composition only if $n = 2, 4, 8$. We combine this with the known results¹⁸ for \mathfrak{F} of characteristic not two to see that in the non-singular case the only possible values of n are $1, 2, 4, 8$.

7. The singular case with $A + A' \neq 0$

Let \mathfrak{F} be a field of characteristic two and $f(x) = xAx'$ such that $E = A + A' \neq 0$ is singular. Then $f(x)$ has the form (12) for $n > 2r > 0$ and we have seen that we may take $\alpha_1, \dots, \alpha_r$ and $\alpha_{2r+1}, \dots, \alpha_n$ all not zero. Thus

$$(75) \quad A = \begin{pmatrix} A_1 & 0 \\ 0 & D \end{pmatrix}, \quad A_1 = \begin{pmatrix} D_1 & I_r \\ 0 & D_2 \end{pmatrix},$$

¹⁸ The method we have used does not require that \mathfrak{F} be algebraically closed and is applicable to obtain a much simpler derivation than that in the literature of the analogous results for \mathfrak{F} of characteristic not two. We take $f(x) = \alpha_1 x_1^2 + \dots + \alpha_n x_n^2$, $g(y) = \beta_1 y_1^2 + \dots + \beta_m y_m^2$ for the α_i and β_j all not zero and in \mathfrak{F} . Then we have seen that we may take $\alpha_1 = \beta_1 = 1$ and have $G_i G'_i = \beta_i I$, $G_i G'_j + G_j G'_i = 0$ for $i \neq j$ and $i, j = 1, \dots, m$. Thus we may take $G_1 = I$ and obtain $G'_i = -G_i$, $G_i^2 = -\beta_i I$, $G_i G_j = -G_j G_i$ for $i \neq j$ and $i, j = 2, \dots, m$. We define μ to be the greatest integer in $\frac{1}{2}m$ and may prove that 2^μ divides n . For $\mathfrak{B}_1 = (I, G_2, G_3, G_2 G_3)$ is a generalized quaternion subalgebra of the total matrix algebra \mathfrak{A} of degree n and $\mathfrak{A} = \mathfrak{B}_1 \times \mathfrak{A}_2$ where \mathfrak{A}_2 is simple of degree n_2 over its center \mathfrak{F} , $n = 2n_2$. If $\mu > 1$ we have $m \geq 6$ and define $G_k G'_k = H_{k-3}$ for $k = 5, \dots, m$ and see that $G_2 H_j - H_j G_2 = G_3 H_j - H_j G_3 = 0$ for $j = 2, \dots, m-3$. Also $H_j^2 = -\beta_j G_{j+3} I$, $H_i H_j = -H_j H_i$ for $i \neq j$ and $i, j = 2, \dots, m-3$. But then the H_i are in \mathfrak{A}_2 and we have the same situation as above with m replaced by $m-3$, n by n_2 , \mathfrak{A} by \mathfrak{A}_2 . It follows that 2 divides n_2 , 2^μ divides n . After μ such steps we obtain the result desired. Moreover if $n = m$ we have shown that 2^μ divides $n = 3\mu, 3\mu+1$, or $3\mu+2$. However if $\mu > 3$ we have $2^\mu > 3\mu+2$, if $\mu = 3$ then $2^\mu = 8$ does not divide $9, 10, 11$, and if $\mu = 2$ then $2^\mu = 4$ only divides $3\mu+2 = 8$. If $\mu = 1$ then $2^\mu = 2$ only divides $3\mu+1 = 4$ and if $\mu = 0$ then $n = 1, 2$.

where D_1 is the non-singular r -rowed diagonal matrix with diagonal elements $\alpha_1 = 1, \alpha_2, \dots, \alpha_r$, D_2 is the r -rowed diagonal matrix with $\alpha_{r+1}, \dots, \alpha_{2r}$ as diagonal elements, D is the $(n - 2r)$ -rowed non-singular diagonal matrix with diagonal elements $\alpha_{2r+1}, \dots, \alpha_n$. We now prove the

LEMMA 5. *The equation (1) is possible only if there exist linear forms u_{2r+1}, \dots, u_n in y_1, \dots, y_m such that*

$$(76) \quad g(y) = \alpha_{2r+1}^{-1}(\alpha_{2r+1}u_{2r+1}^2 + \dots + \alpha_n u_n^2).$$

For proof we observe first that $g(y)[A + A'] = G_y(A + A')G_y'$. Write

$$(77) \quad G_y = \begin{pmatrix} G & H \\ K & L \end{pmatrix},$$

where G is a $2r$ -rowed square matrix and the number of rows and columns in H, K, L are thereby determined. Then

$$(78) \quad A + A' = \begin{pmatrix} C & 0 \\ 0 & 0 \end{pmatrix}, \quad C = \begin{pmatrix} 0 & I_r \\ I_r & 0 \end{pmatrix}.$$

But then C is non-singular and

$$(79) \quad \begin{pmatrix} g(y)C & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} G & H \\ K & L \end{pmatrix} \begin{pmatrix} C & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} G' & K' \\ H' & L' \end{pmatrix} = \begin{pmatrix} GCG' & GCK' \\ KCG' & KCK' \end{pmatrix}.$$

Hence¹⁰ $g(y)C = GCG'$, G is non-singular, $KCG' = 0$, $K = 0$. Also

$$(80) \quad G_y A (G_y')' = G_y \begin{pmatrix} A_1 & 0 \\ 0 & D \end{pmatrix} \begin{pmatrix} G' & 0 \\ H' & L' \end{pmatrix} = \begin{pmatrix} G & H \\ 0 & L \end{pmatrix} \begin{pmatrix} A_1 G' & 0 \\ DH' & DL' \end{pmatrix}.$$

But the diagonal elements of $g(y)A - G_y A (G_y')'$ are all zero and thus the diagonal elements of $g(y)D - LDL'$ are zero. Write $L = (u_{ij})$ for $i, j = 1, \dots, q$, where $q = n - 2r$ and the u_{ij} are linear forms in y_1, \dots, y_m . Compute the element in the first row and column of this matrix and obtain

$$\alpha_{2r+1}g(y) - (\alpha_{2r+1}u_{11}^2 + \dots + \alpha_n u_{1q}^2) = 0.$$

This completes our proof.

A quadratic form $g(y) = yBy'$ is a diagonal form $\beta_1 y_1^2 + \dots + \beta_m y_m^2$ if and only if $B + B' = 0$. Then every quadratic form equivalent to $g(y)$ is also diagonal. Now if $u_i = \sum_{j=1}^m \lambda_{ij} y_j$ with the λ_{ij} in \mathfrak{F} we have $u_i^2 = \lambda_{i1}^2 y_1^2 + \dots + \lambda_{im}^2 y_m^2$. By (76) we have

¹⁰ Observe that then we are led to the study of the same type of equation as in the non-singular case. We show below that $g(y)$ is diagonal and later that there is then an associated field \mathfrak{K} which is purely inseparable of degree 2^f and exponent two over \mathfrak{F} . But then the set of all $2r$ -rowed square matrices contains a subfield generated by the matrices G_i equivalent to \mathfrak{K} and thus 2^f divides $2r$.

LEMMA 6. *The equation (1) is possible for $n > 2r$ only if $g(y)$ is diagonal.*

It follows immediately that if $n > 2r$ for a quadratic form $f(x)$ with $r > 0$ it cannot permit composition.

8. Diagonal quadratic forms

The theory of (1) for $f(x)$ diagonal and \mathfrak{F} of characteristic two is very simple and the results obtainable with little difficulty. We write $f(x)$ as in (11) with the α_i all not zero. Define $\mathfrak{R}(f)$ to be the purely inseparable field $\mathfrak{F}(u_1, \dots, u_n)$, $u_i^2 = \alpha_i$, so that $\mathfrak{R}(f)$ has degree 2^t and exponent one or two over \mathfrak{F} , $n \leq 2^t$. Define

$$(81) \quad u(x) = x_1 u_1 + \dots + x_n u_n,$$

and let $\mathfrak{L}(f)$ be the set of all values $u(\xi)$ for $\xi = (\xi_1, \dots, \xi_n)$ and the ξ_i in \mathfrak{F} . Clearly $\mathfrak{L}(f)$ is a linear subspace of order $\nu \leq n$, $\nu \leq 2^t$, of $\mathfrak{R}(f)$. Moreover $f(x) = [u(x)]^2$.

By permuting the indeterminates x_1, \dots, x_n if necessary we may assume that u_1, \dots, u_ν are linearly independent in \mathfrak{F} and that $u_k = \sum_{j=1}^{\nu} \alpha_{kj} u_j$ for $k = \nu + 1, \dots, n$. Then the linear transformation

$$(82) \quad x_j = w_j - \sum_{k=\nu+1}^n \alpha_{kj} w_k, \quad x_k = w_k \quad (j = 1, \dots, \nu; k = \nu + 1, \dots, n)$$

is non-singular and carries $f(x)$ into $\alpha_1 w_1^2 + \dots + \alpha_\nu w_\nu^2 + (\sum \alpha_{kj}^2 \alpha_j + \alpha_k) w_k^2$. But $u_k^2 = \alpha_k = \sum \alpha_{kj}^2 \alpha_j$. Hence (82) carries $f(x)$ into

$$(83) \quad \alpha_1 w_1^2 + \dots + \alpha_\nu w_\nu^2.$$

We have thus proved

LEMMA 7. *Let $f(x)$ be a diagonal quadratic form and $\mathfrak{L}(f)$ have order ν over \mathfrak{F} . Then $f(x)$ is equivalent to a form $\alpha_1 x_1^2 + \dots + \alpha_\nu x_\nu^2$.*

Note that ν is an invariant of $f(x)$ and that $f(x)$ is not equivalent to a diagonal form in fewer than ν variables. We now prove

LEMMA 8. *Let $f(x) = \alpha_1 x_1^2 + \dots + \alpha_n x_n^2$, $h(x) = \gamma_1 x_1^2 + \dots + \gamma_m x_m^2$ such that n and m are the respective orders of $\mathfrak{L}(f)$, $\mathfrak{L}(h)$. Then there exist linear forms w_1, \dots, w_n in x_1, \dots, x_m such that $h(x) = f(w)$ if and only if $m \leq n$, $\mathfrak{L}(h)$ is contained in $\mathfrak{L}(f)$. Also then $f(x)$ and $h(x)$ are equivalent if and only if $\mathfrak{L}(h) = \mathfrak{L}(f)$.*

For we may construct a field $\mathfrak{F}(u_1, \dots, u_n, v_1, \dots, v_m) = \mathfrak{B}$ such that $u_i^2 = \alpha_i$, $v_j^2 = \gamma_j$. Then \mathfrak{B} contains both $\mathfrak{L}(f)$ and $\mathfrak{L}(h)$, $f(x) = [u(x)]^2$, $h(x) = [v(x)]^2$, where $u(x)$ is defined by (81) and $v(x) = v_1 x_1 + \dots + v_m x_m$. Also $h(x) = f(w)$ implies that $v(x) = u(w)$, every quantity of $\mathfrak{L}(h)$ is in $\mathfrak{L}(f)$, $\mathfrak{L}(f)$ contains $\mathfrak{L}(h)$. Conversely if $\mathfrak{L}(f)$ contains $\mathfrak{L}(h)$ we may take v_1, \dots, v_m as part of a basis of $\mathfrak{L}(f)$ and then replace $f(x)$ by an equivalent form with $\alpha_i = \gamma_i$ for $i = 1, \dots, m$. Then $h(x) = f(w)$ where $w_i = x_i$ for $i = 1, \dots, m$ and $w_k = 0$ for $k = m + 1, \dots, n$. The last part of our lemma is a trivial consequence of what we have already proved.

The result above implies that we may take $n - 2r$ to be the order of the matrix of $\alpha_{2r+1}x_{2r+1}^2 + \cdots + \alpha_n x_n^2$ in Lemma 5 and obtain $m \leq n - 2r$ in that lemma.

We now prove

LEMMA 9. *Let $f(x)$ be a diagonal form $x_1^2 + \alpha_2 x_2^2 + \cdots + \alpha_n x_n^2$ where n is the order of $\mathfrak{F}(f)$. Then $f(x)g(y) = f(z)$ as in (1), (2), if and only if $g(y)$ is equivalent to a diagonal quadratic form $\beta_1 y_1^2 + \cdots + \beta_m y_m^2$ such that $\mathfrak{F}(f)$ contains $\mathfrak{R}(g)$. Moreover n is divisible by the order 2^t of $\mathfrak{R}(g)$ over \mathfrak{F} .*

The result above states that $\mathfrak{F}(f)$ is a linear space of order ν over $\mathfrak{R}(g)$, $n = 2^t \nu$,

$$(84) \quad \mathfrak{F}(f) = \mathfrak{R}(g) + \mathfrak{R}(g)d_2 + \cdots + \mathfrak{R}(g)d_\nu.$$

Then we may take a basis of $\mathfrak{R}(g)$ of order $\sigma = 2^t$ over \mathfrak{F} to be a set of quantities v_1, \cdots, v_σ such that $v_i^2 = \beta_i$ in \mathfrak{F} and also take $d_j^2 = \gamma_j$ for $j = 2, \cdots, \nu$. Then

$$\text{we may take } g(y) = \beta_1 y_1^2 + \cdots + \beta_m y_m^2 \text{ and } f(x) = \sum_{i=1}^{\sigma} \beta_i x_i^2 + \sum_{i=1}^{\sigma} \beta_i \gamma_2 x_{\sigma+i}^2 + \cdots + \sum_{i=1}^{\sigma} \beta_i \gamma_\nu x_{(\nu-1)\sigma+i}^2.$$

To prove this result we note that $f(e) = 1$ and thus $g(y) = f(w)$ for $w = (w_1, \cdots, w_n)$ and the w_i linear forms in y_1, \cdots, y_m . Thus $g(y)$ is a diagonal form. Then $g(y)$ is a diagonal form such that $\mathfrak{F}(g)$ is contained in $\mathfrak{F}(f)$. But then $g(y) = [v(y)]^2$ for $v(y) = v_1 y_1 + \cdots + v_m y_m$ and $v_i^2 = \beta_i$ where the v_i are a basis of $\mathfrak{F}(g)$. If a and b are in $\mathfrak{F}(g)$ the quantity a is in $\mathfrak{F}(f)$. However $[u(x)]^2[v(y)]^2 = [u(z)]^2$, $u(x)v(y) = u(z)$ and ab is in $\mathfrak{F}(f)$. It follows that the product of any two quantities of $\mathfrak{F}(g)$ is in $\mathfrak{F}(f)$, and $\mathfrak{F}(f)$ contains the ring of all polynomials in the quantities of $\mathfrak{F}(g)$. This is the field $\mathfrak{R}(g)$. Conversely if $\mathfrak{F}(f)$ contains $\mathfrak{R}(g)$ we know that $\mathfrak{F}(f) = [\mathfrak{R}(g)]\mathfrak{B}$ where $\mathfrak{B} = (1, d_2, \cdots, d_\nu)$ over \mathfrak{F} and $1, d_2, \cdots, d_\nu$ are a basis of $\mathfrak{F}(f)$ over $\mathfrak{R}(g)$. Then $\mathfrak{F}(g)\mathfrak{R}(g)$ is in $\mathfrak{R}(g)$, $\mathfrak{F}(g)\mathfrak{F}(f)$ is contained in $\mathfrak{F}(f)$. However the equation $g(x)f(y) = f(z)$ is precisely equivalent to the statement that $\mathfrak{F}(g)\mathfrak{F}(f)$ is contained in $\mathfrak{F}(f)$.

If $g(y)$ is equivalent to $f(x)$ we may require that $m = n$ is the order of $\mathfrak{F}(f)$. Then $\alpha_1 f(y) = f(w)$ and $f(y) = \alpha_1^{-1} f(w)$. Lemma 8 then states that $f(x)$ is equivalent to $\alpha_1^{-1} f(x)$. It follows that we may take $\alpha_1 = 1$. Then we may apply Lemma 9 to obtain the result that $\mathfrak{F}(f)$ contains $\mathfrak{R}(f)$. Then $\mathfrak{F}(f) = \mathfrak{R}(f)$. Thus a diagonal quadratic form over a field of characteristic two permits composition if and only if it is the norm form of a purely inseparable field of degree 2^t and exponent two over \mathfrak{F} . We have thus completed the proof of our theorem.

ON INTEGRAL GEOMETRY IN KLEIN SPACES

BY SHIING-SHEN CHERN

(Received September 3, 1940)

The classical results in integral geometry found by Crofton, Poincaré, Cartan, and recently developed by Blaschke¹ and his school, are mostly restricted to Euclidean spaces. It is the object of this paper to give the fundamental concepts of integral geometry in a general space of Klein, by which we mean a number space of n dimensions with a transitive r -parameter group G_r of transformations. The discussion is mainly based on Cartan's theory of Lie's groups.²

The paper is divided into three sections. In §1 we give a brief summary of Cartan's theory of Lie's groups with some results concerning the measures of geometrical elements. In §2 we define the incidence of the geometrical elements of different fields, which plays a fundamental rôle in all subsequent discussions. As applications of the general notions we give in the last section two formulas which are respectively generalizations of the well-known formulas of Crofton and Cauchy.

1. Some Fundamental Notions in Klein Spaces

Let G_r be an abstract r -parameter group of Lie³ with the parameters a^1, \dots, a^r , so that G_r denotes an r -dimensional space and a^1, \dots, a^r are the coordinates in the space. If S_a denotes a point in G_r with the coordinates a^1, \dots, a^r , there are defined in G_r two simply-transitive groups of transformations

$$(1) \quad \begin{aligned} S_a &\rightarrow S_c S_a, \\ S_a &\rightarrow S_a S_c, \end{aligned}$$

called respectively the first and the second groups of parameters. There exists one, and only one, set of r linearly independent Pfaffian forms

$$\omega^1(a, da), \dots, \omega^r(a, da)$$

invariant under the first group of parameters. They are determined up to a linear transformation with constant coefficients and have the property that any

¹ Cf. W. Blaschke, *Vorlesungen über Integralgeometrie*, Bd. I Leipzig 1936; Bd. II Leipzig 1937. These books will be cited respectively as I. G. I and I. G. II.

² Cf. E. Cartan, *La théorie des groupes finis et continus et la géométrie différentielle traitée par la méthode du repère mobile*, Paris 1937. This book will be cited as Cartan, *Théorie des groupes*.

³ For the definition of an abstract group of Lie we may follow that given by Cartan in his book "*La théorie des groupes finis et continus et l'analyse situs*," *Mémoires des Sciences Mathématiques*, Fasc. XLII, Paris 1930.

exterior differential form⁴ of degree p invariant under the first group of parameters is of the form

$$(2) \quad \sum_{i_1, \dots, i_p} A_{i_1, \dots, i_p} [\omega^{i_1} \dots \omega^{i_p}],$$

where A_{i_1, \dots, i_p} are constants. According to Cartan, we call $\omega^1, \dots, \omega^r$ the relative components of G_r . They satisfy the equations of structure of Maurer-Cartan

$$(3) \quad (\omega^i)' = \sum_{j,k=1}^r c_{jk}^i [\omega^j \omega^k], \quad c_{jk}^i + c_{kj}^i = 0, \quad i = 1, \dots, r,$$

where c_{jk}^i are the constants of structure of G_r .

Now let g be a subgroup of G_r . This subgroup g and its left-hand cosets Sg fill up the whole space G_r and have the property that no two of them can coincide without being identical. Thus the varieties Sg are the integral varieties of a completely integrable Pfaffian system. Owing to the invariance of the totality of the cosets Sg with respect to the first group of parameters, the left-hand members of the Pfaffian system are linear combinations of $\omega^1, \dots, \omega^r$ with constant coefficients. We may suppose the system to be

$$(4) \quad \omega^1 = 0, \dots, \omega^n = 0.$$

The theorem of Frobenius then gives

$$(5) \quad c_{jk}^i = 0, \quad i = 1, \dots, n; \quad j, k = n+1, \dots, r,$$

which are the necessary and sufficient conditions for the system (4) to be completely integrable. We say that the subgroup g defines a field of geometrical elements such that each element of the field is represented by a left-hand coset of g or by an integral variety of (4).

We may interpret the geometrical elements defined by (4) as the points of a space E . The first group of parameters becomes then a group of transformations in E , so that, with some necessary assumptions on the analyticity of the equations of the transformations, E is a space in the sense of Klein. It is clear that any space of Klein can be obtained this way. In fact, we need only take G_r to be the group of transformations in the space and g the subgroup of G_r leaving invariant a fixed point.

Let $x^1(a^1, \dots, a^r), \dots, x^n(a^1, \dots, a^r)$ be n independent first integrals of (4), so that dx^1, \dots, dx^n are n linearly independent combinations of $\omega^1, \dots, \omega^n$. Then we have

$$(6) \quad [dx^1 \dots dx^n] = \Delta(a^1, \dots, a^r) [\omega^1 \dots \omega^n], \quad \Delta \neq 0.$$

By the measure of a domain D of elements in E we shall mean an n -tuple integral of the form

$$(7) \quad J = \int_D f(x^1, \dots, x^n) [dx^1 \dots dx^n]$$

⁴ For the notions of exterior differential forms and exterior derivation cf. E. Cartan, *Leçons sur les invariants intégraux*, Chap. VI, VII, Paris 1922.

extended over D such that its value is invariant under the group of transformations in E , i.e., under the first group of parameters. Since this property holds for all domains D , we see from (6) that it is expressed by

$$f\Delta = \text{constant}.$$

Hence a necessary and sufficient condition for the elements of E to possess a measure is that the function $\Delta(a)$ in (6) be a function of x^1, \dots, x^n only. The measure is then defined up to a constant factor and is given by

$$(8) \quad J = \int_D [\omega^1 \dots \omega^n].$$

The condition for the existence of a measure can also be expressed in terms of the constants of structure. Let d and δ be two operations such that d denotes a displacement in the set of elements in E and δ a displacement on an integral variety of (4). The condition for the existence of measure can be written as

$$\delta[\omega^1 \dots \omega^n] = 0.$$

But we have

$$\omega^1(\delta) = \dots = \omega^n(\delta) = 0.$$

The equations of structure (3) then give

$$\delta\omega^i(d) = 2 \sum_{k=1}^n \sum_{j=n+1}^r c_{jk}^i \omega^j(\delta) \omega^k(d), \quad i = 1, \dots, n.$$

From the last equations we get

$$\delta[\omega^1 \dots \omega^n] = 2 \sum_{j=n+1}^r \sum_{i=1}^n c_{ji}^i \omega^j(\delta) [\omega^1 \dots \omega^n].$$

Therefore a necessary and sufficient condition for the elements defined by (4) to possess a measure is

$$(9) \quad \sum_{i=1}^n c_{ji}^i = 0, \quad j = n+1, \dots, r.$$

2. Definition of the Incidence of Elements of Different Fields

Consider two fields of geometrical elements, defined respectively by (4) and

$$(10) \quad w^1 = 0, \dots, w^m = 0,$$

where w^i ($i = 1, \dots, m$) are linear combinations of the relative components with constant coefficients. Each of the systems (4) and (10) being completely integrable, the same must be true of the combined system

$$(11) \quad \omega^1 = 0, \dots, \omega^n = 0, \quad w^1 = 0, \dots, w^m = 0.$$

Of the last system suppose s of them ($s > m, s > n$) be linearly independent, so that there exist $m + n - s$ relations of the form

$$(12) \quad \sum_{i=1}^n b_{\lambda i} \omega^i + \sum_{j=1}^m c_{\lambda j} w^j = 0, \quad \lambda = 1, \dots, m + n - s,$$

where $b_{\lambda i}, c_{\lambda j}$ are constants. The integral varieties of (11) are of dimension $r - s$. Denoting the integral varieties of (4), (10), (11) respectively by $V_{r-n}, V_{r-m}, V_{r-s}$, we see that a V_{r-s} lies completely on a V_{r-n} (or V_{r-m}) if and only if it has a point in common with V_{r-n} (or V_{r-m}). It follows that a V_{r-n} and a V_{r-m} have a V_{r-s} in common if and only if they have a point in common.

By virtue of the above properties we define an element N of (4) and an element M of (10) to be incident when their corresponding integral varieties V_{r-n}, V_{r-m} have a V_{r-s} in common.

Consider the elements M incident with a given element N . In the group space G_r each of the integral varieties V_{r-m} corresponding to M cuts V_{r-n} in a V_{r-s} . The totality of V_{r-s} on a fixed V_{r-n} depends on $s - n$ parameters. Therefore the elements of the field (10) incident with a given element N of (4) depend on $s - n$ parameters. The same property remains naturally true when the two fields are interchanged.

It is important to notice that this definition of incidence includes the ordinary notions of incidence in Euclidean geometry, affine geometry, projective geometry, etc. as particular cases. Take, for instance, the group of motions in the Euclidean plane

$$(13) \quad \begin{cases} x^* = x \cos c - y \sin c + a, \\ y^* = x \sin c + y \cos c + b, \end{cases}$$

where a, b, c are parameters. The relative components are

$$(14) \quad \begin{cases} \omega^1 = \cos c da + \sin c db, \\ \omega^2 = -\sin c da + \cos c db, \\ \omega^3 = dc. \end{cases}$$

As the equations of points and straight lines we may take respectively

$$(15) \quad \omega^1 = 0, \quad \omega^2 = 0$$

and

$$(16) \quad \omega^1 = 0, \quad \omega^3 = 0.$$

In the three-dimensional group space (a, b, c) the elements of the field (15) are represented by lines parallel to the c -axis and those of (16), by the lines

$$\cos c \cdot a + \sin c \cdot b = \text{const.}, \quad c = \text{const.}$$

By interpreting a, b as the Cartesian coordinates in the plane and the equation

$$\cos c \cdot a + \sin c \cdot b = \text{const.}$$

as the equation of line in Hesse's normal form, we see immediately that our definition of incidence coincides with the notion of incidence in the ordinary sense.

Before concluding this section, we want to remark that, instead of using the integral varieties in the group space to characterize the elements of a field, we may employ the geometrically more intuitive idea of families of frames (in French "repère").⁵ But it is sufficient for our purpose to restrict ourselves to one point of view.

3. The Generalized Crofton's Formula and Cauchy's Formula

Let us first consider the geometry in the Euclidean plane. It is well known that the measures for points and lines exist and that they are given respectively by⁶

$$\int [\omega^1 \omega^2] \quad \text{and} \quad \int [\omega^1 \omega^3].$$

Crofton's formula in its simplest form asserts that the measure of the lines incident with the points of a curve is equal to a constant multiple of the length of the curve, each line being counted as many times as the number of its points of intersection with the curve.

To generalize this formula, we must first have the generalized notion of the length of a curve for a p -dimensional variety of points under an arbitrary group of transformations of Lie. For this purpose, take a p -dimensional variety V_p formed by the elements of the field (4). Let u^1, \dots, u^p be the parameters on V_p . Then $\omega^1, \dots, \omega^n$ are linear combinations of du^1, \dots, du^p on V_p . On eliminating du^1, \dots, du^p , we may express $\omega^{p+1}, \dots, \omega^n$ as linear combinations of $\omega^1, \dots, \omega^p$ in the form

$$\omega^\alpha = \sum_{k=1}^p \xi_k^\alpha \omega^k, \quad \alpha = p+1, \dots, n,$$

where ξ_k^α are functions of the parameters a^1, \dots, a^r of the group and of u^1, \dots, u^p . By the method of moving frames of Cartan, it is in general possible (by supposing the variety V_p to be sufficiently general) to determine some or all of the ξ_k^α to be constants or functions of the parameters u^1, \dots, u^p such that the determination is invariant under transformations of the group G_r . If, up to a certain step in the determination of ξ_k^α , the exterior differential forms of degree p

$$[\omega^{i_1} \dots \omega^{i_p}],$$

where i_1, \dots, i_p run over all combinations of $1, \dots, n$, depend on u^1, \dots, u^p only and differ from each other only by constant factors, we say that the variety V_p possesses a p -dimensional area, which is equal to the integral of any one of the

⁵ Cf. Cartan, *Théorie des groupes*, Chap. V, or H. Weyl, *The Classical Groups*, Princeton 1939, pp. 16-17.

⁶ Blaschke, I. G. I, pp. 5-7.

differential forms and is thus determined up to a constant factor. It is of course possible that the area of V_p does not exist, as in the case that V_p is a quadric in projective space. But in the cases which usually occur to us, the p -dimensional area of V_p under G_r exists.

In the case $p = 1$ the p -dimensional area so defined leads to the Pick's invariant of a curve,⁷ which includes the affine arc, projective arc, etc. as particular cases. Pick has proved that *if a group of transformations of order r in the xy -plane transforms transitively the elements of contact of order $r - 2$*

$$x, y, y' = \frac{dy}{dx}, \dots, y^{(r-2)} = \frac{d^{r-2}y}{dx^{r-2}},$$

then a plane curve possesses an intrinsic parameter invariant under the group considered. This intrinsic parameter is called the invariant of Pick. Its existence is an easy consequence of our preceding discussions. In fact, it is only necessary to take $x, y, y', \dots, y^{(r-2)}$ to be parameters of the group and

$$\omega^1 = A^1(x, y, \dots, y^{(r-2)}) dx + B^1(x, y, \dots, y^{(r-2)}) dy = 0,$$

$$\omega^2 = A^2(x, y, \dots, y^{(r-2)}) dx + B^2(x, y, \dots, y^{(r-2)}) dy = 0$$

to be the equations of points. Along a curve we have $dy = y'dx$, so that ω^2 is a constant multiple of ω^1 , and the integral

$$\int \omega^1$$

gives the intrinsic parameter of the curve.

Another important particular case of the notion of p -dimensional area is the case that $p = s - m$ and that V_p consists of all the elements of the field (4) incident with a fixed element of the field (10). In this case, the relations between $\omega^1, \dots, \omega^n$ are obtained from (12) by setting $w^1 = \dots = w^m = 0$ and are therefore

$$\sum_{i=1}^n b_{\lambda i} \omega^i = 0, \quad \lambda = 1, \dots, m + n - s.$$

Since $b_{\lambda i}$ are constants, we may take as the p -dimensional area of this V_p the integral

$$\int [\omega^1 \dots \omega^p].$$

If the value of this integral over V_p is finite, we call it *the measure of the elements of (4) about a fixed element of (10)*.

With these preparations we can give the generalized Crofton's formula in the following theorem:

⁷ Cf., for example, G. Kowalewski, *Allgemeine Natürliche Geometrie und Liesche Transformationsgruppen*, Berlin 1931, pp. 106-110.

Let two fields of elements M, N be defined respectively by the equations (10), (4). Let $p = m + n - s$ and let V_p be a variety of p dimensions formed by the elements of M . If the measure of the elements of N incident with the elements of V_p and the p -dimensional area F of V_p both exist and are finite, then

$$(17) \quad \int [\omega^1 \cdots \omega^n] = cF, \quad c = \text{constant},$$

where the integral is extended over the elements of N incident with the elements of V_p , each element being counted as many times as the number of elements of V_p with which it is incident.

To prove this theorem, consider the $p = m + n - s$ relations (12). Since the relations are independent with respect to $\omega^1, \dots, \omega^n$, we may solve them in terms of $\omega^1, \dots, \omega^p$, obtaining

$$\omega^\lambda = \sum_{j=p+1}^n e_j^\lambda \omega^j + \sum_{k=1}^m f_k^\lambda \omega^k, \quad \lambda = 1, \dots, p,$$

where e_j^λ, f_k^λ are constants. Setting

$$\bar{\omega}^\lambda = \omega^\lambda - \sum_{j=p+1}^n e_j^\lambda \omega^j, \quad \lambda = 1, \dots, m + n - s,$$

and denoting $\bar{\omega}^\lambda$ again by ω^λ , we have

$$(18) \quad \omega^\lambda = \sum_{k=1}^m f_k^\lambda \omega^k, \quad \lambda = 1, \dots, m + n - s.$$

The proof of our theorem then depends on a new form of the expression $[\omega^1 \cdots \omega^n]$. We apply first the method of moving frames of Cartan to V_p . Since by hypothesis the p -dimensional area of V_p exists we can attach to each element of V_p a frame such that within the p -parameter family of frames so attached the exterior differential forms

$$[\omega^{i_1} \cdots \omega^{i_p}]$$

where i_1, \dots, i_p run over all combinations of $1, \dots, m$, differ from each other only by constant factors. In the group space $G_r V_p$ is represented by a p -parameter family of left-handed cosets with respect to a subgroup h . To the frames attached to the elements of V_p there corresponds in G_r a p -dimensional variety W_p such that on every coset corresponding to an element of V_p there lies one and only one point of W_p . Now every integral variety Sh of (10) is cut by the integral varieties of (4) incident to it in the $(r - s)$ -dimensional integral varieties V_{r-s} of (11). The totality of V_{r-s} on Sh depends on $s - m$ parameters. To define a system of coordinates for the V_{r-s} on Sh we may first set up a system of coordinates $\lambda^1, \dots, \lambda^{s-m}$ for the V_{r-s} on h . By the transformation $S_a \rightarrow SS_a$ (S being fixed) the variety h is carried to Sh and we define the coordinates of a V_{r-s} on h to be the coordinates of its image on Sh .

Let u^1, \dots, u^p be the parameters on V_p . As the coordinates of an element

of the field N incident with an element of V_p we may take $u^1, \dots, u^p, \lambda^1, \dots, \lambda^{s-m}$. Now the expression $[\omega^1 \dots \omega^n]$ depends only on the elements of the field N under consideration. This property, called by Blaschke the property of invariance of choice ("Wahlinvarianz"), may be interpreted as follows: Choose on each integral variety of (4) a point and compute the relative components $\omega^1, \dots, \omega^n$ on the n -dimensional variety so obtained. The resulting expression is independent of the choice of these points.

This being clear, we may chose the points on the integral varieties of (4) to be on the V_{r-s} of the integral varieties corresponding to the elements of V_p . Its totality forms a variety of $p + s - m = n$ dimensions, which we call V_n . To find the relative components on V_n notice that the relative components $\omega^i(a, da)$ ($i = 1, \dots, r$) are the parameters of the infinitesimal transformation $S_a^{-1}S_{a+da}$ and that by choosing the infinitesimal transformations $X_i f$ ($i = 1, \dots, r$) such that $X_{m+1}f, \dots, X_r f$ generate the subgroup h leaving invariant a fixed element 0 of the field M , the infinitesimal transformation $S_a^{-1}S_{a+da}$ carrying 0 to a neighboring element P is of the form

$$w^1X_1f + \dots + w^mX_mf + \omega^{m+1}X_{m+1}f + \dots + \omega^rX_rf,$$

where w^1, \dots, w^m are exactly the left-hand members of the system (10).

Let $\theta^1, \dots, \theta^m$ denote the relative components w^1, \dots, w^m on W_p , $\bar{\omega}^{p+1}, \dots, \bar{\omega}^n$ the relative components $\omega^{p+1}, \dots, \omega^n$ on Sh , and let $w^1, \dots, w^m, \bar{\omega}^{p+1}, \dots, \bar{\omega}^n$ denote these components on V_n . As the independent components on V_n we may then take p independent forms among the $\theta^1, \dots, \theta^m$, and $\bar{\omega}^{p+1}, \dots, \bar{\omega}^n$, the latter being Pfaffian forms in $\lambda^1, \dots, \lambda^{s-m}$ only. Now every point of V_n is of the form

$$S_a = S_b R,$$

where S_b and R belong respectively to W_p and h . By writing

$$S_{a+da} = S_{b+db} R^1, \quad R^1 \in h,$$

we get

$$S_a^{-1} S_{a+da} = R^{-1} S_b^{-1} S_{b+db} R \cdot R^{-1} R^1.$$

Since $R^{-1}R^1$ belongs to h , it produces no effect on 0. The relative components w^1, \dots, w^m are thus transformed according to the transformations of the adjoint group which correspond to transformations of h . As R transforms between themselves the infinitesimal transformations of h , we have

$$(19) \quad \begin{cases} w^1 = a_1^1\theta^1 + \dots + a_m^1\theta^m, \\ \vdots \\ w^m = a_1^m\theta^1 + \dots + a_m^m\theta^m. \end{cases}$$

where $a_k^i(i, k = 1, \dots, m)$ are functions of $\lambda^1, \dots, \lambda^{s-m}$.

let u^1, \dots, u^p be the parameters on V_p and let us attach to each element of V_p a frame. Almost all the arguments used above are valid, including the formulas (19), (20). The only difference is that the variety, formally denoted by V_n , is now of $p + s - m$ dimensions. This variety V will be cut by an integral variety V_{r-n} of (4) in a variety $V_{p+s-m-n}$ of $p + s - m - n$ dimensions (supposing $p > m + n - s$). Through each point of $V_{p+s-m-n}$ there passes an integral variety of (10). Its totality consists of all the elements of V_p incident with V_{r-n} . To find the independent components on $V_{p+s-m-n}$ it is only necessary to set $\omega^\lambda = 0$ in (18). The independent Pfaffian forms in the equations so obtained

$$\sum_{k=1}^m f_k^\lambda w^k = 0, \quad \lambda = 1, \dots, m + n - s$$

then give the independent components on $V_{p+s-m-n}$. The generalized Cauchy's formula states that *the integral over all elements of (4) incident with V_p of a $(p + s - m - n)$ -dimensional integral invariant of $V_{p+s-m-n}$ of the form*

$$(23) \quad \int \sum C_{i_1 \dots i_{p+s-m-n}} [w^{i_1} \dots w^{i_{p+s-m-n}}],$$

where $C_{i_1 \dots i_{p+s-m-n}}$ are functions of $\lambda^1, \dots, \lambda^{s-m}$ only, is equal to a constant multiple of the p -dimensional area of V_p , provided that both quantities are finite.

It is only necessary to prove the formula for the case when the sum (23) contains one term. Suppose it be

$$\int C[w^1 \dots w^{p+s-m-n}].$$

By making use of (18), (20), we get

$$C[w^1 \dots w^{p+s-m-n} \omega^1 \dots \omega^n] = F(\lambda^1, \dots, \lambda^{s-m})[\theta^1 \dots \theta^n \bar{\omega}^{m+n-s+1} \dots \bar{\omega}^n],$$

where F is a function of $\lambda^1, \dots, \lambda^{s-m}$. The generalized Cauchy's formula as stated above is then an easy consequence of the relation obtained.

It may be helpful to give an example of the above general discussions. Consider the three-dimensional Euclidean space with its group of motions. Instead of introducing the group space we may take a fixed right-hand rectangular trihedral T_0 in space and take all such trihedrals T as the elements of the space. Then the motions in space and the trihedrals T are in one-to-one correspondence such that to T corresponds the motion carrying T_0 to T . As the parameters of the group we may take the parameters of T . Let P be the origin of T and $\vec{e}_1, \vec{e}_2, \vec{e}_3$ the three unit vectors along the axes of T . Then the relative components of the group are defined by the equations

$$(24) \quad \begin{cases} dP = \omega^1 \vec{e}_1 + \omega^2 \vec{e}_2 + \omega^3 \vec{e}_3, \\ d\vec{e}_i = \omega_i^1 \vec{e}_1 + \omega_i^2 \vec{e}_2 + \omega_i^3 \vec{e}_3, \end{cases} \quad \omega_i^j + \omega = 0, \quad (i, j = 1, 2, 3).$$

The equations of structure of the group are

$$(25) \quad \begin{cases} (\omega^i)' = \sum_{k=1}^3 [\omega^k \omega_k^i], \\ (\omega_k^i)' = \sum_{j=1}^3 [\omega_k^j \omega_j^i], \end{cases} \quad i, k = 1, 2, 3.$$

As the coset in the group space corresponding to a point we take the family of trihedrals having the point as origin. As the coset corresponding to a straight line we take the trihedrals whose origin is on the line and whose vector \vec{e}_3 is along the line. Finally, the coset corresponding to a plane is formed by the trihedrals with origin on the plane and with the third unit vector \vec{e}_3 perpendicular to the plane. By these definitions, the equations of the points, lines, and planes are respectively

$$(26) \quad \begin{cases} \omega^1 = 0, & \omega^2 = 0, & \omega^3 = 0; \\ \omega^1 = 0, & \omega^2 = 0, & \omega_3^1 = 0, & \omega_3^2 = 0; \\ \omega^3 = 0, & \omega_3^1 = 0, & \omega_3^2 = 0. \end{cases}.$$

For all these three fields of elements the measures exist, as may be verified by using the criterion (9). They are then given by the integrals

$$(27) \quad \int [\omega^1 \omega^2 \omega^3], \quad \int [\omega^1 \omega^2 \omega_3^1 \omega_3^2], \quad \int [\omega^3 \omega_3^1 \omega_3^2].$$

Consider now the fields of points and lines and denote them by M, N respectively. Then

$$m = 3, \quad n = 4, \quad s = 5.$$

Put

$$w^1 = \omega^1, \quad w^2 = \omega^2, \quad w^3 = \omega^3.$$

The relations (18) become

$$\omega^1 = w^1, \quad \omega^2 = w^2.$$

To obtain the formula of Crofton in this case take a surface Σ and attach to each point of the surface a trihedral $P\vec{v}_1\vec{v}_2\vec{v}_3$ with origin at this point and with \vec{v}_3 normal to the surface. Choose the trihedral $P\vec{e}_1\vec{e}_2\vec{e}_3$ attached to a line through P such that \vec{e}_1 is on the plane $P\vec{v}_1\vec{v}_2$. Then we have

$$(28) \quad \begin{aligned} \vec{v}_1 &= -\sin \psi \vec{e}_1 - \cos \psi \cos \varphi \vec{e}_2 + \cos \psi \sin \varphi \vec{e}_3, \\ \vec{v}_2 &= \cos \psi \vec{e}_1 - \sin \psi \cos \varphi \vec{e}_2 + \sin \psi \sin \varphi \vec{e}_3, \\ \vec{v}_3 &= \sin \varphi \vec{e}_2 + \cos \varphi \vec{e}_3 \end{aligned}$$

where φ, ψ are the parameters of the lines through P , φ being the angle between \vec{e}_3 and \vec{v}_3 . From

$$dP = \theta^1 \vec{v}_1 + \theta^2 \vec{v}_2 = w^1 \vec{e}_1 + w^2 \vec{e}_2 + w^3 \vec{e}_3,$$

we get

$$(29) \quad \begin{aligned} \omega^1 &= -\theta^1 \sin \psi + \theta^2 \cos \psi, \\ \omega^2 &= -\theta^1 \cos \psi \cos \varphi - \theta^2 \sin \psi \cos \varphi. \end{aligned}$$

It follows that

$$(30) \quad [\omega^1 \omega^2 \omega_3 \omega_3^2] = \cos \varphi [\tilde{\omega}_3^1 \tilde{\omega}_3^2 \theta^1 \theta^2].$$

This is a particular case of the formula (21). The classical proof of Crofton's formula is based on this relation.

Next, take the fields of points and planes and denote them by M, N respectively. Then

$$m = 3, \quad n = 3, \quad s = 5.$$

Put

$$w^1 = \omega^1, \quad w^2 = \omega^2, \quad w^3 = \omega^3.$$

The set of relations (18) consists only of one equation

$$\omega^3 = w^3.$$

As by (29), we get

$$(31) \quad w^3 = \theta^1 \cos \psi \sin \varphi + \theta^2 \sin \psi \sin \varphi.$$

Along the curve of intersection of the plane $P\vec{e}_1\vec{e}_2$ with Σ , we have

$$w^3 = 0$$

and

$$dP = w^1 \vec{e}_1$$

so that \vec{e}_1 is the tangent and w^1 the element of arc. Then we get

$$(32) \quad [w^1 \omega^3 \omega_3^1 \omega_3^2] = [w^1 w^3 \tilde{\omega}_3^1 \tilde{\omega}_3^2] = -\sin \varphi [\theta^1 \theta^2 \tilde{\omega}_3^1 \tilde{\omega}_3^2].$$

From this relation we derive easily Cauchy's formula.

CONSERVATION OF SCHOLARLY JOURNALS

The American Library Association created this last year the Committee on Aid to Libraries in War Areas, headed by John R. Russell, the Librarian of the University of Rochester. The Committee is faced with numerous serious problems and hopes that American scholars and scientists will be of considerable aid in the solution of one of these problems.

One of the most difficult tasks in library reconstruction after the first World War was that of completing foreign institutional sets of American scholarly, scientific, and technical periodicals. The attempt to avoid a duplication of that situation is now the concern of the Committee.

Many sets of journals will be broken by the financial inability of the institutions to renew subscriptions. As far as possible they will be completed from a stock of periodicals being purchased by the Committee. Many more will have been broken through mail difficulties and loss of shipments, while still other sets will have disappeared in the destruction of libraries. The size of the eventual demand is impossible to estimate, but requests received by the Committee already give evidence that it will be enormous.

With an imminent paper shortage attempts are being made to collect old periodicals for pulp. Fearing this possible reduction in the already limited supply of scholarly and scientific journals, the Committee hopes to enlist the cooperation of subscribers to this journal in preventing the sacrifice of this type of material to the pulp demand. It is scarcely necessary to mention the appreciation of foreign institutions and scholars for this activity.

Questions concerning the project or concerning the value of particular periodicals to the project should be directed to Wayne M. Hartwell, Executive Assistant to the Committee on Aid to Libraries in War Areas, Rush Rhees Library, University of Rochester, Rochester, New York. .

THE CLOSURE OPERATORS OF A LATTICE

BY MORGAN WARD

(Received January 29, 1940)

I. INTRODUCTION

1. If \mathfrak{S} is a lattice of elements A, B, \dots , the class of all operators of \mathfrak{S} (that is, one-valued functions $\phi X = \phi(X)$ on \mathfrak{S} to \mathfrak{S}) may be made into a lattice by defining the union δ and cross-cut κ of any set Φ of operators ϕ by¹

$$\delta X = (\dots \phi X \dots), \quad \kappa X = [\dots \phi X \dots], \quad \phi \in \Phi.$$

The union and cross-cut here are taken over all the values ϕX of the operators in Φ for any given X of \mathfrak{S} .

It is easily verified that the operators of \mathfrak{S} form a lattice in which $\phi \supset \psi$ if and only if $\phi X \supset \psi X$ for every X of \mathfrak{S} ; furthermore this lattice is closed, modular, or distributive according as \mathfrak{S} is closed, modular or distributive.²

The operator lattice of a lattice is a concept comparable in generality to the Boolean algebra of all subsets of a lattice. As in the algebra, it is certain distinguished sets of operators which are useful in investigating the given lattice rather than the operator lattice itself.

One obviously important distinguished type is the linear operator. An operator ϕ is said to be linear if for any subset \mathfrak{A} of elements A of \mathfrak{S} , it has one or more of the four properties

$$(1.1) \quad \begin{array}{ll} \text{(i)} \quad \phi(\dots A \dots) = (\dots \phi A \dots), & \text{(iii)} \quad \phi[\dots A \dots] = [\dots \phi A \dots], \\ \text{(ii)} \quad \phi(\dots A \dots) = [\dots \phi A \dots], & \text{(iv)} \quad \phi[\dots A \dots] = (\dots \phi A \dots). \end{array}$$

Here the unions and cross-cuts are taken over all the elements of \mathfrak{A} , and \mathfrak{A} is finite if \mathfrak{S} is not closed. Lattice homomorphisms and homomorphisms with respect to union with properties (i), (iii) and (i) respectively are familiar examples. (Ore 1).

The linear operators and certain associated lattices are important in the study of residuated lattices (Ward-Dilworth 1) as I plan to show in detail elsewhere.³

¹ If \mathfrak{S} is not closed, Φ is assumed to contain only a finite number of operators. A lattice is said to be closed (or "complete" or "continuous") if it contains the union and cross-cut of any subset of elements in it.

² Chain conditions in \mathfrak{S} do not usually carry over to the operator-lattice.

³ The product $\phi\psi$ of two operators ϕ and ψ defined by $\phi\psi X = \phi(\psi(X))$ immediately gives us an associative multiplication over the operator lattice. On the other hand if B is any fixed element of a residuated lattice \mathfrak{S} , the operators μ and ρ defined by $\mu X = BX$, $\rho X = B:X$ have the linear properties $\mu(\dots A \dots) = (\dots \mu A \dots)$, $\rho(\dots A \dots) = [\dots \rho A \dots]$.

I have discussed elsewhere (Ward 2) a type of operator associated with a point lattice,⁴ which may be used to classify all such lattices of finite order.

2. I develop here the properties of a type of operator which is of fundamental importance in the study of certain imbedding problems of ring theory and semi-group theory.⁵ A typical problem of this class is to imbed a system I of elements over which a commutative and associative multiplication is defined in a residuated lattice \mathfrak{S} so as to preserve the multiplication in I and thus to study the arithmetical properties of I . (Clifford 2, Ward-Dilworth 2). The imbedding is effected by defining a suitable type of "ideal" (distinguished subset) of I in the Boolean algebra of its subsets.⁶ A closely related problem is to imbed a semi-ordered set in a closed lattice. (Mac Neille 1).

The "closure operators" introduced here enable us to view all these problems from a unified standpoint, and explain why in all extant theories of ideals as distinguished subsets, the cross-cut of two ideals is the set-theoretic cross-cut of their elements.

II. CLOSURE OPERATORS

3. Let \mathfrak{S} be a closed lattice. An operator ϕ of \mathfrak{S} is said to be a closure operator if it satisfies the following three conditions:⁷

I 1. $A \supset B$ implies that $\phi A \supset \phi B$.

I 2. $\phi \supset \iota$.

I 3. $\phi^2 = \phi$.

Here ι is the identity operator leaving every element of \mathfrak{S} unchanged.

If \mathfrak{T} is any set of elements T of \mathfrak{S} , it may be proved that every closure operator ϕ has the quasi-linear properties

$$(3.1) \quad \phi[\dots \phi T \dots] = [\dots \phi T \dots],$$

$$(3.2) \quad \phi(\dots T \dots) = \phi(\dots \phi T \dots), \quad T \in \mathfrak{T}.$$

No actual linearity is assumed.

THEOREM 3.1. *The cross-cut⁸ of any set of closure operators is again a closure operator.*

⁴ A lattice is called a point lattice if every element in it save the null element is a union of points. Here a point is any element covering the null element. Point lattices include important types of projective geometries, exchange lattices, and Boolean algebras.

⁵ For a discussion of these problems, the reader is referred to Clifford 1, 2 where references are given to the work of Prüfer and others.

⁶ Several definitions are usually possible. See Ward-Dilworth 2.

⁷ These axioms are satisfied by Kuratowski's closure operator over a Boolean algebra with points. (Kuratowski 1). But they are essentially weaker, as Kuratowski's operator is linear with respect to union. Compare also Birkhoff 1.

⁸ In general, no closure properties hold for the union and product of (closure) operators. It may be shown that if ϕ and ψ are operators, then (ϕ, ψ) is an operator if and only if $(\phi, \psi) = \phi\psi = \psi\phi$. Commutativity is thus a necessary condition for the union (ϕ, ψ) to be an operator. It is evidently a sufficient condition for the product $\phi\psi$ to be an operator.

PROOF. Let Φ be a set of closure operators ϕ , and let $\kappa = [\dots \phi \dots]$ be their cross-cut. We shall show that κ satisfies I 1, I 2, I 3.

I 1 is satisfied. For $A \supset B$ implies $\phi A \supset \phi B$ for every $\phi \in \Phi$. Hence $[\dots \phi A \dots] \supset [\dots \phi B \dots]$, $\kappa A \supset \kappa B$. I 2 is satisfied. For since $\phi A \supset A$ for every $\phi \in \Phi$, $[\dots \phi A \dots] \supset A$ or $\kappa \supset \iota$. I 3 is satisfied. For $\kappa^2 A = [\dots \phi \kappa A \dots]$, $\phi \in \Phi$. Now $\phi \supset \kappa$. Hence $\phi A \supset \kappa A$, $\phi^2 A \supset \phi \kappa A$, $\phi A \supset \phi \kappa A$. Accordingly $\kappa \supset \kappa^2$. By I 1 and I 2, $\kappa^2 \supset \kappa$. Hence $\kappa^2 = \kappa$, completing the proof.

4. Let ϕ be a given closure operator, and let $\mathfrak{S}' = \phi \mathfrak{S}$ be the set of all its values $X' = \phi X$ in \mathfrak{S} . By formula (3.1) any subset \mathfrak{X}' of the X' is closed under cross-cut. We may express this fact by writing

$$(4.1) \quad [\dots T' \dots]_{\mathfrak{S}'} = [\dots T' \dots]_{\mathfrak{S}}, \quad T' \in \mathfrak{X}', \quad \mathfrak{X}' \subseteq \mathfrak{S}'.$$

If I is the unit element of \mathfrak{S} , then $I' = I$ divides all elements A' of \mathfrak{S}' . Hence for any subset \mathfrak{L}' of elements L' of \mathfrak{S}' , the class \mathfrak{K}' of all K' such that $K' \supset L'$ is non-empty. We define the union of the L' to be the cross-cut of the K' :

$$(4.2) \quad (\dots L' \dots)_{\mathfrak{S}'} = [\dots K' \dots]_{\mathfrak{S}'}, \quad L' \supset K' \text{ every } L' \text{ of } \mathfrak{L}'.$$

We obtain by a familiar argument:

THEOREM 4.1. *The set \mathfrak{S}' of values of a given closure operator forms a closed lattice within \mathfrak{S} with respect to the operations of union and cross-cut defined by (4.2) and (4.1).*

To each closure operator ϕ we may accordingly assign a lattice $\mathfrak{S}' = \phi \mathfrak{S}$. In particular, $\mathfrak{S} = \iota \mathfrak{S}$. We shall establish a converse result.

Let \mathfrak{S}' now denote a fixed subset of \mathfrak{S} closed under cross-cut and containing the unit element I . We make \mathfrak{S}' into a lattice within \mathfrak{S} by assigning to any subset \mathfrak{L}' of elements of \mathfrak{S}' as in (4.2) a union defined as the cross-cut of the set of all multiples of the elements of \mathfrak{L}' .

We next define an operator ϕ on \mathfrak{S} to \mathfrak{S}' as follows: If A is any element of \mathfrak{S} , then ϕA is the cross-cut of all elements B' of \mathfrak{S}' such that $B' \supset A$. Then ϕ is a closure operator, for I 1, I 2, I 3 are evidently satisfied. Furthermore, $\phi \mathfrak{S} = \mathfrak{S}'$.

We have thus established a one-to-one correspondence between the closure operators of \mathfrak{S} and subsets of \mathfrak{S} closed under cross-cut and containing I . The lattice $\mathfrak{S}' = \phi \mathfrak{S}$ and the operator ϕ will be said to belong to one another.

It is also easily proved from formula (3.2) that \mathfrak{S}' is a sublattice of \mathfrak{S} if and only if the closure operator belonging to \mathfrak{S}' is linear with respect to union.

THEOREM 4.2. *The closure operators of any closed lattice themselves form a lattice within the operator lattice of \mathfrak{S} .*

PROOF. Let Σ denote the set of all closure operators of \mathfrak{S} . By formula (3.1), the cross-cut of any set of such operators is again a closure operator. Furthermore the operator ω defined by $\omega A = I$, every A of \mathfrak{S} , is obviously a closure operator dividing every other closure operator. Hence we may define the union of any set Φ of such operators as the cross-cut of the non-empty set of closure operators containing every operator of Φ .

We may evidently define lattice operations on the set of all subsets \mathfrak{S}' of \mathfrak{S} closed under cross-cut and containing I by the rules

$$(4.3) \quad \begin{aligned} [\dots \mathfrak{S}' \dots] &= [\dots \phi \dots] \mathfrak{S}, & \phi \in \Phi, & \quad \mathfrak{S}' = \phi \mathfrak{S} \subset \Phi \mathfrak{S} \\ (\dots \mathfrak{S}' \dots) &= (\dots \phi \dots) \mathfrak{S}. \end{aligned}$$

The lattices \mathfrak{S}' thus form a lattice simply isomorphic with the lattice Σ of closure operators. We shall return to these operations at the close of the next section.

5. Consider an operator ϕ belonging to a set consisting of two elements I and T of \mathfrak{S} . It follows from the previous theorems that ϕ is characterized by

$$(5.1) \quad \phi A = I \text{ if } T \nabla A, \quad \phi A = T \text{ if } T \supset A, \quad A \text{ any element of } \mathfrak{S}.$$

We call ϕ the two-valued operator belonging to T . Since $\phi \mathfrak{S}$ is a sub-lattice of \mathfrak{S} , ϕ is linear with respect to union, as is directly evident from (5.1). It is easy to prove

THEOREM 5.1. *The ideal operator ϕ belonging to any set \mathfrak{S}' of elements of \mathfrak{S} which is closed under cross-cut and contains I is the operator cross-cut of all the two-valued operators belonging to elements of \mathfrak{S}' .*

If \mathfrak{I} is any set of elements T of \mathfrak{S} containing I , we obtain a lattice \mathfrak{S}' within \mathfrak{S} containing \mathfrak{I} by adjoining to \mathfrak{I} the cross-cuts of all sets of its elements. \mathfrak{S}' is evidently the smallest such lattice containing \mathfrak{I} . We call \mathfrak{S}' the imbedding lattice of \mathfrak{I} , and its corresponding closure operator the "imbedding operator" of \mathfrak{I} . We shall use the letter θ to denote an imbedding operator.

THEOREM 5.2. *If θ is the imbedding operator of a set \mathfrak{I} of elements of \mathfrak{S} containing I , then the value of θ for any element A of \mathfrak{S} is given by the formula*

$$(5.2) \quad \theta A = [\dots T \dots], \quad T \supset A, \quad T \in \mathfrak{I}.$$

PROOF. We have $\theta A = S'$ where S' lies in $\mathfrak{S}' = \theta \mathfrak{S}$. Hence S' is the cross-cut of a certain set of the T in \mathfrak{I} . Now since $\theta A \supset A$, every such T divides A . But since $\theta T = T$ if $T \in \mathfrak{I}$, $T \supset A$ implies that $T \supset \theta A = S'$. Hence (5.2) follows.

THEOREM 5.3. *Let ϕ and ψ be any two closure operators of \mathfrak{S} . Then $\phi \supset \psi$ if and only if the lattice belonging to ψ contains the lattice belonging to ϕ in the set-theoretic sense.*

PROOF. Assume that $\phi \supset \psi$ and let $A \in \phi \mathfrak{S}$. Then $\phi A = A$. By I 1, $\phi A \supset \psi A$. Hence $A \supset \psi A$. Therefore by I 2, $A \equiv \psi A$ or $A \in \psi \mathfrak{S}$. Since ϕ and ψ are the imbedding operators of their respective lattices, the converse follows from Theorem 5.2.

The following corollaries are immediate:

COROLLARY 5.31. *Let θ be the imbedding operator belonging to any set \mathfrak{I} of elements of \mathfrak{S} containing I , and let ψ be any closure operator such that ψ leaves every element of \mathfrak{I} invariant. Then ψ divides θ .*

COROLLARY 5.32. *The imbedding operator of any set is the union of all closure operators which leave every element of the set invariant.*

COROLLARY 5.33. *The union operation on the lattices which belong to closure operators defined by (4.3) is the operation of taking the set-theoretic cross-cut of their elements.*

It is this correspondence between operator union and set-theoretic cross-cut which makes the ideal operators of importance in imbedding problems.

III. APPLICATIONS TO IMBEDDING PROBLEMS

6. Let I be a set of elements a, b, \dots semi-ordered with respect to a division relation $x | y$ and containing a unit element 1 dividing every other element. The following problem has been considered by Mac Neille: (Mac Neille 1). To construct a closed lattice \mathfrak{S}' such that: (i) \mathfrak{S}' contains a subset of elements A', B', \dots which may be set in a one-to-one correspondence $x \leftrightarrow X'$ with a, b, \dots ; (ii) If $a \leftrightarrow A'$ and $b \leftrightarrow B'$, then

$$(6.1) \quad a | b \text{ in } I \text{ implies } A' \supset B' \text{ in } \mathfrak{S}'.$$

$$(6.2) \quad A' \supset B' \text{ in } \mathfrak{S}' \text{ implies } a | b \text{ in } I.$$

We call such a construction an "isomorphic imbedding" of the set I . If we do not require (6.2), we speak of a "homomorphic imbedding" of I .

We shall solve these problems by determining suitable ideal operators in the lattice \mathfrak{B} (Boolean algebra) of all subsets of I . In other words, we shall determine all ideal operators ϕ of \mathfrak{B} such that $\mathfrak{S}' = \phi\mathfrak{B}$ will be a suitable lattice.

Consider first the condition (6.1). Let $T = (t)$ be a subset of I consisting of the single element t . Since $T' = \phi T \supset T$, we must have $t \in \phi T$. But by (6.1), if $t | y$ in I , $\phi T \supset \phi(y)$. Hence if $t | y$, $y \in \phi T$. Thus (6.1) implies that ϕT must contain all elements y of I such that $t | y$.

For a homomorphic imbedding, no further conditions are imposed on the values of ϕT . But if the imbedding is isomorphic and $\phi T \supset \phi(X)$, then (6.2) requires that $t | x$. Hence ϕT must consist only of elements x of I such that $t | x$.

We let \mathfrak{T} denote the set of all $T' = \phi(t)$, $t \in I$ for any ideal operator ϕ . We call the elements of \mathfrak{T} the principal ideals of I .

It is evident from the preceding section that any ideal operator of \mathfrak{B} leaving every element of \mathfrak{T} invariant will solve our initial imbedding problem, and that the simplest of these operators is the imbedding operator of the set \mathfrak{T} itself; for its lattice $\theta\mathfrak{B}$ is the smallest lattice in the set-theoretic sense in which the imbedding can be made in \mathfrak{B} . The isomorphism between I and \mathfrak{T} with respect to division shows that this same minimal property of $\theta\mathfrak{B}$ will apply to any isomorphic imbedding of I in any closed lattice \mathfrak{S}' whatever; within the lattice \mathfrak{S}' there must lie a lattice simply isomorphic to $\theta\mathfrak{B}$. $\theta\mathfrak{B}$ is the lattice defined in Mac Neille 1 by "Dedekind cuts."

A similar situation occurs for homomorphic imbeddings. For a homomorphic imbedding, the "principal ideals" A', B', \dots which make up the set \mathfrak{T} are not

uniquely determined by the corresponding elements a, b, \dots of I ; for if $a \leftrightarrow A'$, A' may contain elements of I not divisible by a . But once the set \mathfrak{T} of principal ideals is chosen, the imbedding operator of I gives the smallest lattice in which the particular homomorphic imbedding can be performed.

7. If A is any subset of I , let A be the subset of all elements l such that $l \mid k$ for every k in A , and let A' be the subset of all elements a such that $l \mid a$ for every l in A . Then the operator

$$(7.1) \quad A' = \theta A$$

is the isomorphic imbedding operator of the set I discussed above. This result follows easily from Theorem 5.3. For a detailed discussion, the reader may consult Ward-Dilworth 2 or Clifford 1, to whom this definition of θ is originally due.⁹

CALIFORNIA INSTITUTE OF TECHNOLOGY.

REFERENCES

- | | |
|----------------------------|--|
| GARRETT BIRKHOFF | 1 Duke Math. Journal, 3 (1931) pp. 443-454. |
| A. H. CLIFFORD | 1 Bull. Am. Math. Soc. 40 (1934) pp. 326-330. |
| | 2 These Annals (2) 39 (1938) pp. 594-610. |
| C. KURATOWSKI | 1 <i>Topologie</i> 1, Warsaw (1933). |
| H. M. MAC NEILLE | 1 Trans. Am. Math. Soc. 42 (1937) pp. 416-460. |
| O. ORE | 1 These Annals, (2) 36 (1935) pp. 406-437. |
| M. WARD AND R. P. DILWORTH | 1 Trans. Am. Math. Soc. vol. 45 (1939), pp. 335-354. |
| M. WARD | 2 Unpublished. |

⁹ The identity of this operator and Mac Neille's operator was pointed out to me by Dr. A. H. Clifford in a letter. The definition (7.1) is used in Ward-Dilworth 2 to imbed any ovum (semi-group) in a residuated lattice of ideals.

UNSTABLE MINIMAL SURFACES WITH SEVERAL BOUNDARIES¹

By MAX SHIFFMAN

(Received August 21, 1940; revised October 1, 1941)

TABLE OF CONTENTS

INTRODUCTION.....	197
PART I. DEGENERATE DOMAINS. THE SPACES \mathfrak{R}_k AND \mathfrak{P} . THE FUNCTIONAL $E[\mathfrak{f}]$	
1. The Domains of Representation.....	198
2. A Metric for \mathfrak{R}_2	201
3. The Metric Space \mathfrak{R}_k	204
4. The Connectivity Numbers of \mathfrak{R}_k	206
5. The Space \mathfrak{P}	207
6. The Continuity of the Functional $E[\mathfrak{f}]$	208
PART II. APPLICATION TO UNSTABLE MINIMAL SURFACES	
7. Linear Paths of Surfaces.....	211
8. The Reducibility Condition.....	213
9. The Case of Polygonal Boundaries.....	215
10. The Variational Condition.....	217
11. The Main Theorems. Remarks.....	221
BIBLIOGRAPHY.....	222

INTRODUCTION

In recent years, the question of unstable minimal surfaces bounded by a single contour has been attacked with success.² It has been shown that the Morse critical point theory applies to minimal surfaces bounded by the contour Γ (provided Γ satisfies certain restrictions). The contributions to the theory have been made by the author [11], [12], by Morse and Tompkins [7], [8] and by Courant [3].

In the present paper we shall show how to extend the theory to cover minimal surfaces of genus zero bounded by an arbitrary number k of non-intersecting contours $\Gamma_1, \Gamma_2, \dots, \Gamma_k$.³ In a general but not precise way, the main result may be stated thus: if the Morse theory can be shown to apply to each of the contours $\Gamma_1, \dots, \Gamma_k$ individually, it applies to all the contours $\Gamma_1, \dots, \Gamma_k$ together.

¹ Presented to the American Mathematical Society, April 26, 1940.

² Concerning the Plateau problem and minimal surfaces in general, see [1], [2], [4], [5], [10]. Numbers in brackets refer to the bibliography at the end.

³ In the meantime a paper by Morse and Tompkins [9] has appeared which considers the case of two boundaries.

A first step in the development of the theory is the following (see [12], [7]). Let \mathfrak{P} denote the space of admissible potential surfaces $\mathfrak{x}(u, v)$ with a finite Dirichlet integral $D[\mathfrak{x}]$, where $D[\mathfrak{x}] = \frac{1}{2} \iint (\mathfrak{x}_u^2 + \mathfrak{x}_v^2) du dv$; and let \mathfrak{P}_N denote the subspace of \mathfrak{P} for which $D[\mathfrak{x}] \leq N$. The first step, used even in the minimum theory, is that \mathfrak{P}_N be compact.

For the case of several boundaries, this requirement is *not* satisfied for ordinary potential surfaces. To retain the compactness of \mathfrak{P}_N it is necessary to introduce degenerate potential surfaces, consisting of several pieces, and their corresponding degenerate parameter domains. These degenerate domains and surfaces have been so selected that the discontinuities of the Dirichlet integral can be reduced to its discontinuities for single boundaries. More precisely, suppose that \mathfrak{x} is an admissible potential surface defined over a domain G having k boundaries. Let G_μ be the region of the plane bounded by the μ^{th} boundary curve C_μ and intersecting G , and $\bar{\mathfrak{x}}_\mu$ the potential surface defined over G_μ with boundary values identical to \mathfrak{x} on C_μ . Set

$$E[\mathfrak{x}] = D_G[\mathfrak{x}] - \sum_{\mu=1}^k D_{G_\mu}[\bar{\mathfrak{x}}_\mu].$$

Then the degenerate domains and surfaces introduced here are such that $E[\mathfrak{x}]$ is a *continuous* functional.

These two theorems, the compactness of \mathfrak{P}_N and the continuity of $E[\mathfrak{x}]$, form the backbone of this paper. In Part II, the continuity of $E[\mathfrak{x}]$ yields very simply the reducibility property of $D[\mathfrak{x}]$ provided that reducibility is known for simple contours.

For two boundaries, the theory can be developed much more simply. But the decisive step is the case $k > 2$.

PART I. DEGENERATE DOMAINS. THE SPACES \mathfrak{R}_k AND \mathfrak{P} . THE FUNCTIONAL $E[\mathfrak{x}]$

1. The Domains of Representation

Let $\Gamma_1, \Gamma_2, \dots, \Gamma_k$ be k non-intersecting closed Jordan curves in space. On each Γ_μ select three distinct points P_μ, Q_μ, R_μ . We shall consider potential surfaces $\mathfrak{x}(u, v)$, defined over certain normal domains of the $w = u + iv$ plane, which map the boundaries of these domains continuously and monotonically into $\Gamma_1, \Gamma_2, \dots, \Gamma_k$.

Choose, for the ordinary (i.e., non-degenerate) domains of representation in the w -plane, circular regions G consisting of an annular ring, unit circle as outer boundary, with $k - 2$ additional circular holes. Let C_μ denote the circular boundary of G mapped by $\mathfrak{x}(u, v)$ into Γ_μ , and let p_μ, q_μ, r_μ be three distinct points of C_μ mapped by $\mathfrak{x}(u, v)$ into P_μ, Q_μ, R_μ respectively. These specified points $p_\mu, q_\mu, r_\mu, \mu = 1, 2, \dots, k$, will be considered as part of the domain of representation G . In considering the class of admissible potential surfaces

$\xi(u, v)$, not only do the boundary values of $\xi(u, v)$ vary but also the domains of representation (including the points p_μ, q_μ, r_μ).

It is possible to normalize G , by performing linear transformations and reflections, in such a way that the following conditions are satisfied:

- (a) C_1 is the unit circle;
- (b) C_2 is concentric with the unit circle;
- (c) p_1, q_1, r_1 lie in counterclockwise order around C_1 ; and
- (d) for $k = 2$, p_1 is at the point $w = 1$ of the w -plane, while for $k > 2$ the center of C_k is on the positive real axis.

We shall suppose throughout that G has been normalized in this way.

It is necessary to define the limit of a set of domains, and for compactness to introduce degenerate domains. Our procedure is motivated by considering a sequence $\xi^n(u, v)$ of admissible potential surfaces with uniformly bounded Dirichlet functional $D[\xi^n] \leq N$. Let G^n be the ordinary domains over which the ξ^n are defined. Then obtain a limit domain G^∞ and a limit potential surface ξ^∞ such that all pieces which contribute a non-vanishing term to the Dirichlet integral are retained.

The only possibilities for the degeneracy of the sequence ξ^n are:⁴ the circles of G^n degenerate as $n \rightarrow \infty$; and the boundary values of ξ^n are not equicontinuous. Now, the non-equicontinuity of the boundary values of ξ^n on C_μ is equivalent to the assertion that the minimum distance between the three points $p_\mu^n, q_\mu^n, r_\mu^n$ has zero as a limiting value as $n \rightarrow \infty$. Since p_μ, q_μ, r_μ are considered as part of the domain G , all degeneracies have been transferred to the domain. This limits the discussion of the sequence ξ^n , so far as degeneration is concerned, to that of the domains G^n .

Accordingly, let G^n be a sequence of ordinary domains, and (by choosing a subsequence if necessary) suppose that $p_\mu^n, q_\mu^n, r_\mu^n$ converge to definite points of the w -plane as $n \rightarrow \infty$. The sequence has no limit domain if any pair of circles with radii above a positive bound approach each other as $n \rightarrow \infty$ (this would contradict $D[\xi^n] \leq N$), and we suppose that this is not the case. The G^n 's are said to *tend to degeneracy* if the $3k$ limit points of $p_\mu^n, q_\mu^n, r_\mu^n$ as $n \rightarrow \infty$ are not all distinct.

If the domains G^n do not tend to degeneracy as $n \rightarrow \infty$, the *limit* domain G^∞ is immediate: G^∞ is the circular domain whose points $p_\mu^\infty, q_\mu^\infty, r_\mu^\infty$ and circles C^∞ are the limits of those of G^n .

If the sequence G^n does tend to degeneracy, one or several of the three following types of degeneration must occur (for convenience a circle of G^n with radius approaching 0 as $n \rightarrow \infty$ is called a *small* circle; one whose radius does not approach 0, a *large* circle): (a) One or more small circles tend to a point A away from all other circles; (b) One or more small circles and a large circle C^n approach a point A , while at most one of the points $p_\mu^n, q_\mu^n, r_\mu^n$ on C^n approach A as $n \rightarrow \infty$; (c) At least two of the points $p_\mu^n, q_\mu^n, r_\mu^n$ on a large circle C^n

⁴ For all the properties of the sequence ξ^n used in the next two pages, see [1].

approach a point A . In cases (b) and (c) let A^n be a point on the large circle C^n approaching A ; for (a), let A^n be A .

In any of these three situations, describe in G^n a circle or arc of circle K^n with A^n as center and with the radius $(\rho^n)^{\frac{1}{2}}$, where ρ^n is the maximum distance from A^n to the small circles and points p^n, q^n, r^n which may be approaching A . (It can be shown that the oscillation of x^n on K^n tends to 0.) K^n splits G^n into two regions \tilde{G}_1^n and \tilde{G}_2^n , large and small respectively. Also, in cases (b) and (c), K^n divides C^n into two arcs α^n and β^n , where α^n contains at least two of the points p^n, q^n, r^n and β^n at most one of these points. The arc α^n is to be counted in the limit as the circle C ; and β^n in the limit as coordinated to a single point of C , (since the oscillation of x^n on β^n approaches 0 as $n \rightarrow \infty$). Separate \tilde{G}_1^n and \tilde{G}_2^n and normalize them both by performing linear transformations and inversions. In this normalization, disregard the K^n ; temporarily, \tilde{G}_1^n and \tilde{G}_2^n are considered as bounded by circles.

Continue as above with the regions \tilde{G}_1^n and \tilde{G}_2^n separately, but with the following provisos: K^n is to be disregarded, and β^n is to be disregarded if it approaches a point away from all other circles. One finally obtains the decomposition of G^n (a subsequence of the original G^n 's) into several regions $H_1^n, H_2^n, \dots, H_l^n$ each of which degenerates no further. A passage to the limit yields a degenerate domain G^∞ consisting of several distinct circular regions H_1, H_2, \dots, H_l ; G^∞ is called the *limit* of this final subsequence G^n . (A subsequence of the x^n can then be chosen to converge uniformly to a limit potential surface x^∞ defined over the various regions which compose G^∞ .)

We are now in a position to completely define degenerate domains. A *degenerate* domain G consists of several distinct circular regions, each supposed normalized. In all the regions together there are exactly k circles of type C_μ , $\mu = 1, 2, \dots, k$, called *boundary* circles, and on each C_μ three distinct points p_μ, q_μ, r_μ ; there may be additional circles, denoted by c_μ, c'_μ, \dots and called *point* circles, which are coordinated to single points on C_μ likewise marked c_μ, c'_μ, \dots . (The limit potential surface x^∞ would take constant values on each point circle c_μ equal to its value at the corresponding point c_μ on C_μ .) The domain is of genus zero, i.e., if the point circles c_μ, c'_μ, \dots are attached to the corresponding points c_μ, c'_μ, \dots , then every closed curve disconnects the resulting surface. Finally there is no region with a single boundary and that boundary a point circle c_μ (in such a region $x(u, v)$ would be identically constant).

A limit of a sequence of degenerate domains is obtained as previously, considering each region separately and taking the circles and points c_μ, c'_μ, \dots into account.

Finally, a domain G is a limit of a set X of domains, degenerate or ordinary, if X contains a sequence having G as a limit in the above sense. The totality of domains, ordinary and degenerate, together with this definition of limit will be denoted by \mathfrak{R}_k . \mathfrak{R}_k seems to be the natural totality of domains of representation for potential functions.

Figures 1 and 2 illustrate the various kinds of domains with two boundaries.

The degenerate domain in fig. 2(a) is the limit of a sequence of ordinary domains (fig. 1) whose boundary values are not equicontinuous on C_2 ; fig. 2(b) is the limit of a sequence of ordinary domains whose boundary values are not equicontinuous on C_1 ; fig. 2(c), whose boundary values not equicontinuous on both C_1 and C_2 ; and fig. 2(d), for which the radius of C_2 approaches zero. For more than two boundaries the degenerate domains become much more numerous and complicated.

It is necessary to know that \mathfrak{R}_k is a topologic space, and more, that it is metric. This will be established by imbedding \mathfrak{R}_k into an Euclidean space of sufficiently high dimension. We begin with the case of two boundaries.

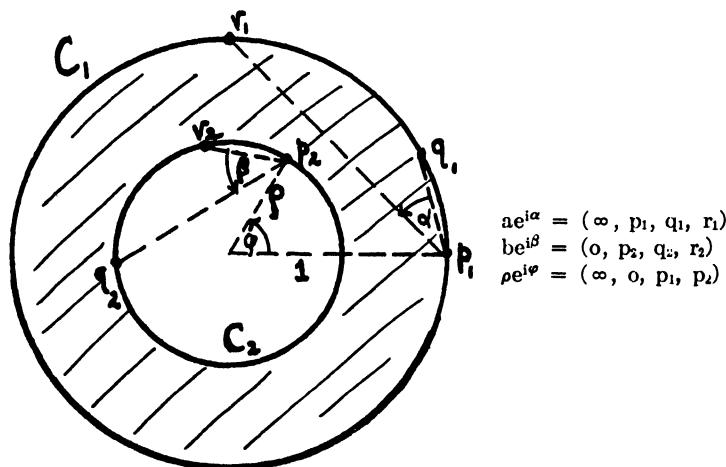


FIGURE 1. Ordinary domains in \mathfrak{R}_2

2. A Metric for \mathfrak{R}_2

Consider first the ordinary domains in \mathfrak{R}_2 (the case of two boundaries) as in fig. 1. The domains depend on 6 variable parameters which we will choose below so as to vary continuously with the domain. The principal difficulty in this choice is caused by the degenerate domains. In the passage to a limit from a sequence of ordinary domains to a degenerate domain, many linear transformations and reflections were performed; and the degenerate domains may possibly contain points c_μ corresponding to point circles c_μ . The parameters must be chosen to take these two possibilities into account.

The following six parameters $a, \alpha, b, \beta, \rho, \varphi$ are selected (angles are measured in the interval from $-\pi$ to π , excluding $-\pi$):

$$(1) \quad \begin{cases} (\infty, p_1, q_1, r_1) = ae^{i\alpha} & (0 < a < \infty, 0 < \alpha < \pi), \\ (0, p_2, q_2, r_2) = be^{i\beta} & (0 < b < \infty, \beta \neq 0, \pi), \\ (\infty, 0, p_1, p_2) = \rho e^{i\varphi} & (0 < \rho < 1), \end{cases}$$

where (w_1, w_2, w_3, w_4) means the cross ratio $\frac{w_1 - w_3}{w_1 - w_4} \cdot \frac{w_2 - w_4}{w_2 - w_3}$. Cross ratio is chosen so as to obtain invariance under linear transformations, and the

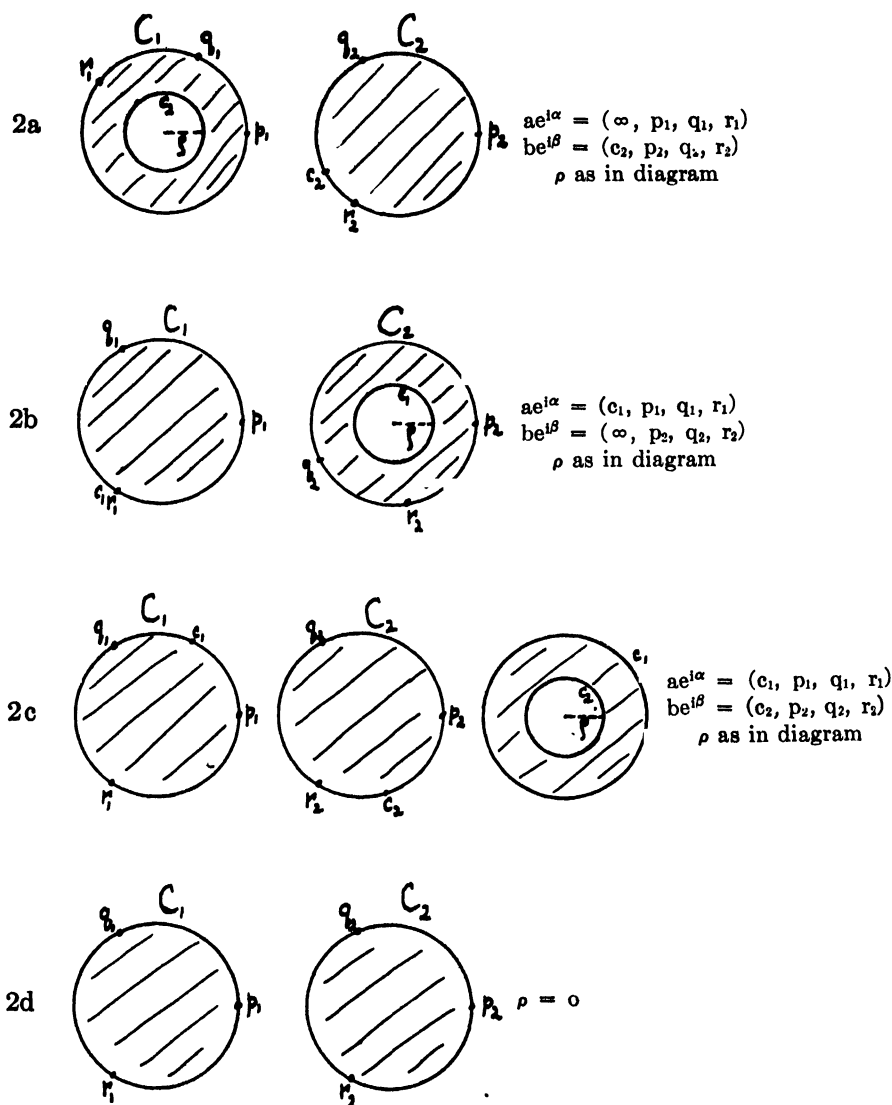


FIGURE 2. Degenerate domains in \mathcal{R}_2

points $\infty, 0$ correspond to points of type c_1, c_2 in degenerate domains. The precise reason for the choice of the points $\infty, 0$ in the cross ratio expressions is made apparent in the proof of Lemma 1 below, part (b).

The geometric meanings of $\alpha, \beta, \rho, \varphi$ are indicated in fig. 1, while $a = \frac{\overline{p_1 r_1}}{p_1 q_1}$ and $b = \frac{\overline{p_2 r_2}}{p_2 q_2}$. It is easily established geometrically that the parameters $a, \alpha, b, \beta, \rho, \varphi$ subject to the inequalities in (1) (these inequalities are necessary for the non-degeneracy and normalization of the domain) uniquely determine the domain.

For degenerate domains, the parameters $a, \alpha, b, \beta, \rho, \varphi$ are defined as in fig. 2. The undefined parameters may have any value (e.g., in fig. 2(a), φ may have any value). For a degenerate domain, note that either $ae^{i\alpha}$ is real (so that $\alpha = 0$ or π , or $a = 0$ or ∞), or $be^{i\beta}$ is real, or $\rho = 0$.

A remaining difficulty is the indeterminateness of some of the parameters in the case of degenerate domains. A method for eliminating this difficulty is to introduce a set of new parameters which, when defined in terms of $a, \alpha, b, \beta, \rho, \varphi$, contain factors vanishing for degenerate domains. This of course necessitates introducing more than 6 new parameters. One such possibility is the set of nine parameters $x, \nu = 1, \dots, 9$, interpreted as the coordinates of a point x in Euclidean space \mathfrak{E}_9 , defined by

$$(2) \quad \begin{cases} x_1 = \frac{\rho}{1-\rho}, & x_2 = \rho \frac{a}{1+a}, \\ x_3 = \rho \frac{b}{1+b}, & x_4 + ix_5 = \rho \frac{a}{1+a^2} e^{i\alpha}, \\ x_6 + ix_7 = \rho \frac{b}{1+b^2} \left(e^{i|\beta|} + \rho \sin \alpha \frac{a}{1+a^2} e^{i\beta} \right), \\ x_8 + ix_9 = \rho \sin \alpha \sin \beta \cdot \frac{a}{1+a^2} \cdot \frac{b}{1+b^2} \cdot e^{i\varphi}. \end{cases}$$

If $a = \infty, a/(1+a)$ and $a/(1+a^2)$ mean 1 and 0 respectively; and similarly for b . The complicated expression for $x_6 + ix_7$ is necessary because reflections may occur in passing from a sequence of ordinary domains to a degenerate domain; this is made clear in the proof of Lemma 1, part (b).

A simple examination shows that a unique point in \mathfrak{E}_9 corresponds to a single domain in \mathfrak{R}_2 , degenerate or not; and to different domains in \mathfrak{R}_2 correspond different points in \mathfrak{E}_9 . The set of points in \mathfrak{E}_9 corresponding to the whole of \mathfrak{R}_2 is easily shown to form a closed 6-dimensional set extending to infinity (in the x_1 -direction).

LEMMA 1. *The one-to-one mapping of \mathfrak{R}_2 onto a point set of \mathfrak{E}_9 , given by (1), fig. 2, and (2), is a topologic mapping.*

PROOF. It is required to show that the one-to-one correspondence preserves the limit relation. Let G^n be a sequence of domains of \mathfrak{R}_2 and x^n the corresponding points in \mathfrak{E}_9 . Since $\rho^n \rightarrow 1$ implies $x_1^n \rightarrow \infty$, and conversely, the sequence G^n has no limit domain if and only if x^n has no limit point. It remains

to establish that $x^n \rightarrow x$ whenever $G^n \rightarrow G$, where x is the point in \mathfrak{E} , corresponding to the domain G . This is trivial if the domains G^n do not degenerate as $n \rightarrow \infty$. If the sequence G^n tends to degeneracy, the possibilities are (see §1):

(a) $\rho^n \rightarrow 0$. By §1, the limit domain G is given in fig. 2(d), and x is therefore the origin of \mathfrak{E} . The equations (2) show that $x^n \rightarrow x$.

(b) Two of the points p_1^n, q_1^n, r_1^n converge to the same point A of the unit circle, while p_2^n, q_2^n, r_2^n converge to distinct points. Let B be the point on the unit circle diametrically opposite to A . The limit domain G is obtained by describing K^n , separating \tilde{G}_1^n and \tilde{G}_2^n , and normalizing each of these regions. Hence G is of the type shown in fig. 2(b). The normalization of \tilde{G}_2^n is performed by inverting with respect to K^n , transforming and inverting so that p_1^n, q_1^n, r_1^n go into the specified points $\theta_1, \theta_2, \theta_3$. The point c_1 in H_2 is the limit of the resulting point B in \tilde{G}_2^n . Since cross ratio is invariant under linear transformations,

$$(B, p_1^n, q_1^n, r_1^n) \rightarrow (c_1, p_1, q_1, r_1).$$

By noting that

$$(B, p_1^n, q_1^n, r_1^n) = \frac{B - q_1^n}{B - r_1^n} \cdot \frac{p_1^n - r_1^n}{p_1^n - q_1^n} \quad \text{and} \quad (\infty, p_1^n, q_1^n, r_1^n) = \frac{p_1^n - r_1^n}{p_1^n - q_1^n},$$

it is easy to show that (B, p_1^n, q_1^n, r_1^n) and $(\infty, p_1^n, q_1^n, r_1^n)$ approach the same value. Thus,

$$(\infty, p_1^n, q_1^n, r_1^n) \rightarrow (c_1, p_1, q_1, r_1), \quad \text{or} \quad a^n e^{i\alpha^n} \rightarrow a e^{i\alpha}$$

(where $\alpha = 0$ or π , or $a = 0$ or ∞).

\tilde{G}_1^n is normalized by inverting with respect to C_2 , expanding, rotating and possibly reflecting. Hence $(0, p_2^n, q_2^n, r_2^n)$ approaches either (∞, p_2, q_2, r_2) or its conjugate complex, so that $b^n e^{i|\beta^n|} \rightarrow b e^{i|\beta|}$. Since clearly $\rho^n \rightarrow \rho$, an examination of equations (2) shows that $x^n \rightarrow x$.

(c) A similar argument applies if two of p_2^n, q_2^n, r_2^n approach the same point A' on C_2 , while p_1^n, q_1^n, r_1^n approach distinct points. The limit domain G is of the type in fig. 2(a).

(d) Similarly, if two of p_1^n, q_1^n, r_1^n approach A , and two of p_2^n, q_2^n, r_2^n approach A' . The limit domain G is given in fig. 2(c).

Like results are obtained if the G^n are degenerate.

In all cases therefore, $G^n \rightarrow G$ implies that $x^n \rightarrow x$. The lemma is established.

Thus, \mathfrak{R}_2 is a metric space. The metric could be defined as follows: the distance between two domains of \mathfrak{R}_2 is the Euclidean distance between their corresponding points in \mathfrak{E} .

3. The Metric Space \mathfrak{R}_k

We return to the case of k boundaries and construct a metric for \mathfrak{R}_k . The parameters which will be used for any domain G of \mathfrak{R}_k are the 9 parameters of the preceding section for each pair of circular boundaries of G . Accordingly,

suppose first that G is an ordinary domain of \mathfrak{R}_k , and C_μ, C_ν ($\mu < \nu$) any pair of boundaries of G with the specified points $p_\mu, q_\mu, r_\mu, p_\nu, q_\nu, r_\nu$. Nine coordinates $x_1^{\mu\nu}, \dots, x_9^{\mu\nu}$ may be associated to this pair of boundaries. They are determined by equations (2) of §2, where the parameters $a, \alpha, b, \beta, \rho, \varphi$ are given by

$$(3) \quad \begin{cases} ae^{i\epsilon\alpha} = (s_\mu, p_\mu, q_\mu, r_\mu), \\ be^{i\epsilon\beta} = (s_\nu, p_\nu, q_\nu, r_\nu), \\ \rho e^{i\epsilon\varphi} = (s_\mu, s_\nu, p_\mu, p_\nu). \end{cases}$$

Here s_μ and s_ν are the two points which are mutually inverse with respect to both circles C_μ and C_ν , and s_μ is on the side of C_μ opposite to C_ν ; also $0 < \alpha < \pi$ and $\epsilon = \pm 1$. These parameters and coordinates are invariant under linear transformations and reflections. They are the invariant generalizations of the equations (1) of §2.

When G is a degenerate domain, the six parameters $a, \alpha, b, \beta, \rho, \varphi$ for the pair C_μ, C_ν ($\mu < \nu$) of boundaries are defined as invariant generalizations of figs. 1, 2(a)–2(d). In this generalization, s_μ and s_ν replace ∞ or 0, and ϵ times an angle (where $\epsilon = \pm 1$) replaces the angle. Thus,

1) if C_μ and C_ν occur in the same region, the parameters are given by equations (3) (see fig. 1);

2) if C_μ and a point circle c_ν are in the same region,

$$\left. \begin{aligned} ae^{i\epsilon\alpha} &= (s_\mu, p_\mu, q_\mu, r_\mu), \\ be^{i\epsilon\beta} &= (c_\nu, p_\nu, q_\nu, r_\nu), \\ \rho &= |(s_\mu, s_\nu, p_\mu, p'_\nu)|, \end{aligned} \right\} \quad (\text{see fig. 2(a)})$$

where p'_ν is any point on the circle c_ν ;

3) if a point circle c_μ and a boundary circle C_ν are in the same region, the parameters are the invariant generalizations of fig. 2(b). (Cf. 2) above.)

4) if point circles c_μ and c_ν are in the same region, the parameters are the generalizations of fig. 2(c). (Cf. 2) above.)

5) if neither C_μ nor any c_μ occurs in the same region as C_ν or any c_ν , then $\rho = 0$ (see fig. 2(d)). Since the domain is of genus zero exactly one of the above occurs for a given pair μ, ν .

The parameters $a, \alpha, b, \beta, \rho, \varphi$ for the pair C_μ, C_ν of boundaries of G determine the 9 coordinates $x_1^{\mu\nu}, \dots, x_9^{\mu\nu}$ by equations (2). We therefore have, for any domain G in \mathfrak{R}_k , a unique point x in the $9k(k-1)/2$ Euclidean space \mathfrak{E} with the coordinates $x_1^{\mu\nu}, \dots, x_9^{\mu\nu}$ for all $\mu, \nu = 1, 2, \dots, k$, and $\mu < \nu$. A simple discussion shows that to different domains in \mathfrak{R}_k correspond different points in \mathfrak{E} .

LEMMA 2. *The one-to-one correspondence given above of \mathfrak{R}_k onto a point set of \mathfrak{E} is a topologic correspondence.*

PROOF. The lemma follows from Lemma 1 if one notes that essentially the process in §1 treats two particular boundaries C_μ, C_ν exactly as they would be treated if they were alone.

Thus, \mathfrak{R}_k is a metric space. Furthermore, the discussion in §1 yields the result that the set of points in \mathfrak{E} corresponding to the whole of \mathfrak{R}_k is a closed set extending to infinity. The fact that the $6k - 6$ dimensional open set (open in \mathfrak{R}_k) of all ordinary domains is everywhere dense in \mathfrak{R}_k completes the proof of

THEOREM 1. *\mathfrak{R}_k is a $6k - 6$ dimensional metric space in which bounded sets are compact.*

4. The Connectivity Numbers of \mathfrak{R}_k

The connectivity numbers of \mathfrak{R}_k will be in the sense of Vietoris theory, considering only cycles and homologies which lie in bounded subsets of \mathfrak{R}_k . The Vietoris character of the cycles will be used in an essential way in the proof of the following theorem.

THEOREM 2. *The connectivity numbers, $R_0, R_1, \dots, R_n, \dots$, of \mathfrak{R}_k have the values:*

$$R_0 = 1, \quad R_1 = R_2 = \dots = R_n = \dots = 0.$$

PROOF. The essential idea of the proof is to deform the whole space \mathfrak{R}_k into a single point, namely the domain consisting of k unit circles (which is the most degenerate domain).

(a) Denote the invariant ρ corresponding to the boundaries C_μ, C_ν ($\mu < \nu$) by $\rho^{\mu\nu}$, and set $\sigma = \text{minimum}_{\mu, \nu} \rho^{\mu\nu}$. Designate the set of all domains of \mathfrak{R}_k for which $\sigma \leq \delta$ by F_δ . We shall first deform \mathfrak{R}_k into F_δ , where δ is any positive number.

The deformed domain G_t , $0 \leq t \leq 1$, will be obtained from the domain G by contracting each circle, except the unit circle, by a certain factor. It is important in this that all the G 's be normalized the following way: in each region of G the order for determining which circle shall be the unit circle, the concentric circle, etc. is C_1 or c_1, C_2 or c_2, \dots, C_k or c_k . If G belongs to F_δ , set $G_t = G$. If the σ for G is $> \delta$, let $\eta (< 1)$ be that contraction of the circles of G which takes G into a domain G_1 for which $\sigma = \delta$; define G_t as the domain obtained from G by contracting its circles by the factor $(1 - t) + t\eta$.

It is necessary to show that G_t is continuous in G and t , i.e., $G_t^n \rightarrow G_t$ whenever $G^n \rightarrow G, t^n \rightarrow t$. This is trivial if $\sigma^n \leq \delta$; if $\sigma^n > \delta$, it follows by noting that G^n can degenerate further only by the coalescence of two of $p_\mu^n, q_\mu^n, r_\mu^n$ whereas contraction to form G_t^n retains the relative position of these points.

One easily notices by consulting equations (2) that the above deformation deforms any bounded set B of \mathfrak{R}_k over a bounded set B' into F_δ ; i.e., for a given bounded set B there is another bounded set B' independent of δ such that B is deformed within B' into F_δ .

(b) Let z^m be any Vietoris m -cycle on a bounded set B . Because of (a), $z^m \sim w_\delta^m$, where w_δ^m is a Vietoris m -cycle on $B' \cdot F_\delta$ and the homology takes place over B' . Since $B' \cdot F_\delta$ is arbitrarily close, for δ sufficiently small, to the closed compact set $B' \cdot F_0$, it follows from a basic theorem for Vietoris cycles that $z^m \sim w_0^m$, where w_0^m is a Vietoris cycle on $B' \cdot F_0$.⁵

⁵ A proof of an essentially equivalent theorem is contained in p. 20-23, Theorem 5.1, of [6].

(c) F_0 consists of the domains G for which at least one of $\rho^{\mu\nu}$ is zero; set $\sigma' =$ minimum of the remaining $\rho^{\mu\nu}$. Designate the set of all domains of F_0 for which $\sigma' \leq \delta$ by $F_{0,\delta}$. By contracting circles, as in (a), F_0 can be deformed into $F_{0,\delta}$ for any positive δ . As in (b), $z^m \sim w_0^m \sim w_{0,0}^m$ where $w_{0,0}^m$ is a Vietoris m -cycle on $B'' \cdot F_{0,0}$.

Continue with $F_{0,0}$ as with F_0 , etc. One finally obtains $z^m \sim w^m$, where w^m is a Vietoris m -cycle on the set F consisting of those domains G of \mathfrak{R}_k for which all the $\rho^{\mu\nu}$ are zero. But F contains a single point, namely the domain consisting of k unit circles, bounded by C_1, C_2, \dots, C_k respectively, with the points p_μ, q_μ, r_μ at the specified places $\theta_1, \theta_2, \theta_3$ of C_μ . The theorem follows immediately.

5. The Space \mathfrak{P}

We are now prepared to discuss the space \mathfrak{P} of potential surfaces bounded by $\Gamma_1, \dots, \Gamma_k$. An element of \mathfrak{P} is a potential surface $\mathfrak{x}(u, v)$ defined over a domain G of \mathfrak{R}_k and satisfying the following conditions: $\mathfrak{x}(u, v)$ maps each C_μ continuously and monotonically onto Γ_μ , and p_μ, q_μ, r_μ into the prescribed points P_μ, Q_μ, R_μ of Γ_μ ; $\mathfrak{x}(u, v)$ takes constant values at each point circle c_μ , equal to its value at the corresponding point c_μ on C_μ ; and $D_G[\mathfrak{x}]$ is finite, where $D_G[\mathfrak{x}]$ is the sum of the Dirichlet integrals of \mathfrak{x} , $\frac{1}{2} \iint (\mathfrak{x}_u^2 + \mathfrak{x}_v^2) du dv$, over the various regions which compose G .

A surface \mathfrak{x} in \mathfrak{P} is determined by the domain G over which it is defined and its boundary values on each of the k circles C_μ . Since the points p_μ, q_μ, r_μ have already been considered part of the domain G , the boundary values of \mathfrak{x} will be defined in the following way. Map the circle C_μ onto a unit circle by a linear transformation or reflection so that p_μ, q_μ, r_μ go into $\theta_1, \theta_2, \theta_3$ respectively. The resulting values of \mathfrak{x} on this circle will be called the *boundary values* of \mathfrak{x} on C , and denoted by $\mathfrak{x}_\mu(\theta)$. The *distance* between two elements \mathfrak{x} and \mathfrak{y} of \mathfrak{P} can now be defined by

$$|\mathfrak{x} - \mathfrak{y}| = |G - H| + \sum_{\mu=1}^k \max_{0 \leq \theta \leq 2\pi} |\mathfrak{x}_\mu(\theta) - \mathfrak{y}_\mu(\theta)|$$

where G and H are the domains of \mathfrak{x} and \mathfrak{y} , and $|G - H|$ is the distance between G, H in \mathfrak{R}_k .

The subspace of all elements \mathfrak{x} of \mathfrak{P} for which $D[\mathfrak{x}] \leq N$ will be denoted by \mathfrak{P}_N . To avoid confusion with the case of a single boundary Γ , the latter space will be designated by \mathfrak{P}^Γ ; the complete notation for the present space \mathfrak{P} is then $\mathfrak{P}^{\Gamma_1, \Gamma_2, \dots, \Gamma_k}$.

It will be convenient to use the following notation. A potential surface in \mathfrak{P} is to be designated by \mathfrak{x} or $\mathfrak{x}(u, v)$, its boundary values in the above sense by $\mathfrak{x}_\mu(\theta)$, and the potential surface defined over the unit circle with the boundary values $\mathfrak{x}_\mu(\theta)$ by \mathfrak{x}_μ . The part of the plane interior or exterior to C_μ , according as G is interior or exterior to C_μ , is denoted by G_μ ; the potential surface defined

over G_μ , with the same values on C_μ as $\mathfrak{z}(u, v)$, by $\bar{\mathfrak{z}}_\mu$ (thus $\bar{\mathfrak{z}}_\mu$ is the linear transform of \mathfrak{z}_μ back to G_μ).

THEOREM 3. $\mathfrak{P} = \mathfrak{R}_k \times \mathfrak{P}^{\Gamma_1} \times \mathfrak{P}^{\Gamma_2} \times \cdots \times \mathfrak{P}^{\Gamma_k}$, where $\cdot \times \cdot$ signifies the topologic product.⁶

PROOF. Let \mathfrak{z} be any potential surface with domain G and boundary values $\mathfrak{z}_\mu(\theta)$. Let K_μ be a circle in G concentric to C_μ and near it; let A_μ be the annular ring between C_μ and K_μ ; and set $G' = G - \sum_{\mu=1}^k A_\mu$. We have

$$D_G[\mathfrak{z}] = \sum_{\mu=1}^k D_{A_\mu}[\mathfrak{z}] + D_{G'}[\mathfrak{z}].$$

Obviously $D_{G'}[\mathfrak{z}] < \infty$, so that $D_G[\mathfrak{z}] < \infty$ if and only if $D_{A_\mu}[\mathfrak{z}] < \infty$ for every μ . In A_μ , set $\mathfrak{z}' = \mathfrak{z} - \bar{\mathfrak{z}}_\mu$. Since \mathfrak{z}' has bounded derivatives near K_μ and is zero on C_μ , $D_{A_\mu}[\mathfrak{z}'] < \infty$. Hence, by the triangle inequality, $D_{A_\mu}[\mathfrak{z}] < \infty$ if and only if $D_{A_\mu}[\bar{\mathfrak{z}}_\mu] < \infty$; and $D_{A_\mu}[\bar{\mathfrak{z}}_\mu] < \infty$ if and only if $D_{G_\mu}[\bar{\mathfrak{z}}_\mu] < \infty$. The theorem follows by noting that $D_{G_\mu}[\bar{\mathfrak{z}}_\mu] < \infty$ is equivalent to: \mathfrak{z}_μ lies in $\mathfrak{P}^{\Gamma_\mu}$.

THEOREM 4. \mathfrak{P}_N is compact and closed for every N .

PROOF. Let $\mathfrak{z}^n(u, v)$, with domains G^n , be a sequence of surfaces in \mathfrak{P}_N . There is a subsequence of the G^n 's converging to G^∞ . Transform each domain G^n linearly so that a given C_ν is the unit circle and the points p_ν, q_ν, r_ν are at the specified places $\theta_1, \theta_2, \theta_3$. As in [1], the boundary values $\mathfrak{z}_\nu^n(\theta)$ are equicontinuous. A final subsequence, written $\mathfrak{z}^n(u, v)$, is obtained such that $\mathfrak{z}^n \rightarrow \mathfrak{z}^\infty$. By referring to the geometric definition of convergence given in §1, it follows that the Dirichlet functional is lower semicontinuous. Hence $D[\mathfrak{z}^\infty] \leq N$, and the theorem is proved.

6. The Continuity of the Functional $E[\mathfrak{z}]$

Let \mathfrak{z} be a potential surface with continuous boundary values $\mathfrak{z}_\mu(\theta)$, and indicate the Dirichlet integral of \mathfrak{z}_μ taken over the unit circle by $D_0[\mathfrak{z}_\mu]$. Define the functional $E[\mathfrak{z}]$ by

$$E[\mathfrak{z}] = D[\mathfrak{z}] - \sum_{\mu=1}^k D_0[\mathfrak{z}_\mu].$$

In this section, we shall prove that $E[\mathfrak{z}]$ is continuous in \mathfrak{P} . This will reduce the behavior of $D[\mathfrak{z}]$ to that of $D_0[\mathfrak{z}_\mu]$.

Let G be the domain of the potential surface \mathfrak{z} , and G_μ that region of the plane which is bounded by C_μ and intersects G . Draw a circle K_μ in G sufficiently near C_μ , $\mu = 1, \dots, k$. Let G' be the subdomain of G bounded by the circles K_μ and any point circles c_μ , and G'_μ that subdomain of G_μ bounded by the circle K_μ . Finally, let $\bar{\mathfrak{z}}_\mu$ be the potential surface defined over G_μ and having the boundary values of \mathfrak{z} on C_μ ; \mathfrak{z}_μ is the linear transform of $\bar{\mathfrak{z}}_\mu$ to a unit circle.

⁶ Furthermore, for any number N there is a compact subset R of \mathfrak{R}_k and a number M , while for any R and M there is an N' , such that $\mathfrak{P}_N \subset R \times \mathfrak{P}_M^{\Gamma_1} \times \cdots \times \mathfrak{P}_M^{\Gamma_k} \subset \mathfrak{P}_{N'}$. This follows from Theorem 5 to be proved below.

We have

$$E[\xi] = D[\xi] - \sum_{\mu=1}^k D_{\sigma_\mu}[\bar{\xi}_\mu] = \lim_{K_\mu \rightarrow C_\mu} \{D_{\sigma'}[\xi] - \sum_{\mu} D_{\sigma'_\mu}[\bar{\xi}_\mu]\};$$

and

$$\begin{aligned} D_{\sigma'}[\xi] - \sum_{\mu} D_{\sigma'_\mu}[\bar{\xi}_\mu] &= \sum_{\mu} \int_{K_\mu} \xi \frac{\partial \xi}{\partial n} ds + \sum_{c_\mu} \int_{c_\mu} \xi \frac{\partial \xi}{\partial n} ds - \sum_{\mu} \int_{K_\mu} \bar{\xi}_\mu \frac{\partial \bar{\xi}_\mu}{\partial n} ds \\ &= \sum_{\mu} \int_{K_\mu} (\xi - \bar{\xi}_\mu) \frac{\partial \bar{\xi}_\mu}{\partial n} ds + \sum_{\mu} \int_{K_\mu} \xi \frac{\partial (\xi - \bar{\xi}_\mu)}{\partial n} ds + \sum_{c_\mu} \int_{c_\mu} \xi \frac{\partial \xi}{\partial n} ds. \end{aligned}$$

If $K_\mu \rightarrow C_\mu$, the first integral on the right hand side approaches zero since $\xi - \bar{\xi}_\mu = 0$ on C_μ . Hence

$$(4) \quad E[\xi] = \sum_{\mu=1}^k \int_{c_\mu} \xi \frac{\partial (\xi - \bar{\xi}_\mu)}{\partial n} ds + \sum_{c_\mu} \int_{c_\mu} \xi \frac{\partial \xi}{\partial n} ds.$$

This is an expression for $E[\xi]$ in terms of boundary integrals.

THEOREM 5. *Let G^n be a sequence of domains in \mathcal{R}_k converging to G , and ξ^n potential surfaces over G^n converging uniformly⁷ to the potential surface ξ over G . Then*

$$E[\xi^n] \rightarrow E[\xi].$$

PROOF. First consider the case when the G^n are ordinary domains in \mathcal{R}_k . If the limit domain G is also ordinary, the theorem follows from (4). For, $\xi^n \rightarrow \xi$ and $\xi^n - \bar{\xi}_\mu^n \rightarrow \xi - \bar{\xi}_\mu$ uniformly near C_μ . Since $\xi^n - \bar{\xi}_\mu^n = 0$ on C_μ^n , it can be extended by reflection across C_μ^n and $\partial(\xi^n - \bar{\xi}_\mu^n)/\partial n$ converges uniformly to $\partial(\xi - \bar{\xi}_\mu)/\partial n$ on C_μ . The second term on the right hand side of (4) is not involved since we have assumed that G^n and G are ordinary domains.

Suppose that the limit domain G is degenerate. The continuity of $E[\xi]$ will be established by induction on the number k of boundaries. We shall distinguish two cases: a) the origin is enclosed by a large circle or is approached by a large circle; and b) all the circles approaching the origin are small circles.

(a) Enclose this large circle and the set of circles approaching this large circle by a curve K containing no other circles. Let G_1^n be the domain bounded by all the circles outside K , and G_2^n the domain bounded by all the circles inside K (G_1^n is a bounded domain, G_2^n unbounded). Let the limits of G^n , G_1^n , G_2^n , considered merely as regions of the plane, be G^* , G_1^* , G_2^* respectively. Define η^n and δ^n over the domains G_1^n and G_2^n respectively as the bounded potential surfaces with boundary values equal to those of ξ^n . Denote the limits of ξ^n , η^n , δ^n by ξ^* , η^* , δ^* ; they are potential surfaces over G^* , G_1^* , G_2^* . We shall first show that $D_{\sigma^n}[\xi^n] - D_{\sigma_1^n}[\eta^n] - D_{\sigma_2^n}[\delta^n] \rightarrow D_{\sigma^*}[\xi^*] - D_{\sigma_1^*}[\eta^*] - D_{\sigma_2^*}[\delta^*]$.

Designate the parts of G^n and G^* exterior to K by A^n and A^* , and interior

⁷ This means that the boundary values of ξ^n in the sense of §5, page 207 converge uniformly to the corresponding boundary values of ξ .

to K by B^n and B^* . Let K_{in} and K_{ex} be the interior and exterior of K . From $G^n = A^n + B^n$, $G_1^n = A^n + K_{\text{in}}$, $G_2^n = B^n + K_{\text{ex}}$, one obtains

$$D_{G^n}[\xi^n] - D_{G_1^n}[\eta^n] - D_{G_2^n}[\delta^n] = \{D_{A^n}[\xi^n] - D_{A^n}[\eta^n]\} - D_{K_{\text{in}}}[\eta^n] \\ + \{D_{B^n}[\xi^n] - D_{B^n}[\delta^n]\} - D_{K_{\text{ex}}}[\delta^n];$$

and a similar decomposition for $D_{G^*}[\xi^*] - D_{G_1^*}[\eta^*] - D_{G_2^*}[\delta^*]$. Now,

$$D_{A^n}[\xi^n] - D_{A^n}[\eta^n] = D_{A^n}[\xi^n - \eta^n, \xi^n + \eta^n] \\ = \int_K (\xi^n - \eta^n) \frac{\partial(\xi^n + \eta^n)}{\partial n} ds \rightarrow \int_K (\xi^* - \eta^*) \frac{\partial(\xi^* + \eta^*)}{\partial n} ds \\ = D_{A^*}[\xi^*] - D_{A^*}[\eta^*];$$

$D_{K_{\text{in}}}[\eta^n] \rightarrow D_{K_{\text{in}}}[\eta^*]$; and like results for $D_{B^n}[\xi^n] - D_{B^n}[\delta^n]$ and $D_{K_{\text{ex}}}[\delta^n]$. Hence

$$(5) \quad D_{G^n}[\xi^n] - D_{G_1^n}[\eta^n] - D_{G_2^n}[\delta^n] \rightarrow D_{G^*}[\xi^*] - D_{G_1^*}[\eta^*] - D_{G_2^*}[\delta^*].$$

The domains G_1^n and G_2^n have k_1 and k_2 boundaries respectively, where $k_1 + k_2 = k$, and the theorem is supposed true for domains with less than k boundaries. Let G^n , G_1^n , G_2^n when considered as domains of \Re_k , \Re_{k_1} , \Re_{k_2} have G , G_1 , G_2 as their limit domains respectively. G_1 consists of G_1^* and other regions $H_{(1)}$; G_2 consists of G_2^* and other regions $H_{(2)}$; and G consists of G^* and the regions $H_{(1)}$ and $H_{(2)}$. Indicate the boundary curves of G_1^n by C_{μ_1} , and those of G_2^n by C_{μ_2} . By the induction,

$$(6) \quad D_{G_1^n}[\eta^n] - \sum_{\mu_1} D_0[\xi_{\mu_1}] \rightarrow D_{G_1^*}[\eta^*] + D_{H_{(1)}}[\xi] - \sum_{\mu_1} D_0[\xi_{\mu_1}]$$

and

$$(7) \quad D_{G_2^n}[\delta^n] - \sum_{\mu_2} D_0[\xi_{\mu_2}] \rightarrow D_{G_2^*}[\delta^*] + D_{H_{(2)}}[\xi] - \sum_{\mu_2} D_0[\xi_{\mu_2}].$$

Adding (5), (6) and (7), one obtains the desired result

$$(8) \quad D_{G^n}[\xi^n] - \sum_{\mu} D_0[\xi_{\mu}] \rightarrow D_{G^*}[\xi^*] + D_{H_{(1)}}[\xi] + D_{H_{(2)}}[\xi] - \sum_{\mu} D_0[\xi_{\mu}] \\ = D_G[\xi] - \sum_{\mu} D_0[\xi_{\mu}].$$

(b) In this case, when all the circles approaching the origin are small circles, the procedure is the same as the above. The only modification is that the region G_2^* is the whole plane, δ^* is identically constant, and $\eta^* = \xi^*$. The relation (5) reads $D_{G^n}[\xi^n] - D_{G_1^n}[\eta^n] - D_{G_2^n}[\delta^n] \rightarrow 0$, and the final result (8) is the same.

There remains the case when the G^n are themselves degenerate. One may suppose, without loss of generality, that all the G^n have the same type of degeneracy: G^n consists of the distinct regions G_1^n, \dots, G_l^n where the G_i^n have the same boundary and point circles for all n . The above proof then applies to each sequence G_i^n separately. The theorem is completely established.

Theorem 5 shows that the discontinuities of the functional $D[\xi]$ are due to the boundary values of ξ and not to the domains, and occur in the same manner as each $D_0[\xi_\mu]$. It should be noted that $E[\xi]$ can be so defined, by (4), so as to exist even if $D[\xi]$ is infinite. Theorem 5 will still apply.

If ξ and η are potential surfaces having the same domain G , define the cross E -functional $E[\xi, \eta]$ by

$$E[\xi, \eta] = D[\xi, \eta] - \sum_{\mu=1}^k D_0[\xi_\mu, \eta_\mu].$$

The bilinear formula applies:

$$E[\xi + \eta] = E[\xi] + 2E[\xi, \eta] + E[\eta].$$

Consequently theorem 5 has as a corollary

THEOREM 6. *Let ξ^n, η^n be two potential surfaces both defined over the domain G^n of \mathcal{R}_k , where $n = 1, 2, \dots$. Let the sequences ξ^n, η^n converge uniformly⁷ to the potential surfaces ξ, η defined over the domain G . Then*

$$E[\xi^n, \eta^n] \rightarrow E[\xi, \eta].$$

PART II. APPLICATION TO UNSTABLE MINIMAL SURFACES

7. Linear Paths of Surfaces

To obtain the Morse relations for minimal surfaces bounded by $\Gamma_1, \Gamma_2, \dots, \Gamma_k$, it is necessary to establish a variational condition and a reducibility condition. The latter requires discussing special paths of surfaces in \mathfrak{P} .

Let us suppose that each Γ_μ is rectifiable, and select the representation $g_\mu(\theta)$, $0 \leq \theta \leq 2\pi$, of Γ_μ in which the parameter θ is proportional to the arc length on Γ_μ measured from the point P_μ and in the direction $P_\mu Q_\mu R_\mu$. Each $g_\mu(\theta)$ has the property

$$(9) \quad \left| \frac{g_\mu(\theta') - g_\mu(\theta'')}{\theta' - \theta''} \right| \leq \frac{L}{2\pi}$$

where L is the largest of the lengths of $\Gamma_1, \dots, \Gamma_k$. Any other representation $\xi_\mu(\theta)$ of Γ_μ is given by

$$\xi_\mu(\theta) = g_\mu(\lambda_\mu(\theta))$$

where $\lambda_\mu(\theta)$ is a monotonic and continuous function of θ . Call $\lambda_\mu(\theta)$ the monotonic function determined by $\xi_\mu(\theta)$.

Let ξ_0 and ξ_1 be any two potential surfaces in \mathfrak{P} having the same domain G , and with the boundary values (in the sense defined in §5, part I) $\xi_\mu(\theta; 0)$ and $\xi_\mu(\theta; 1)$ respectively. Define $\xi(t)$, $0 \leq t \leq 1$, as the potential surface with the same domain G and boundary values $\xi_\mu(\theta; t)$ given by

$$(10) \quad \xi_\mu(\theta; t) = g_\mu[(1-t)\lambda_\mu(\theta; 0) + t\lambda_\mu(\theta; 1)]$$

where $\lambda_\mu(\theta; 0)$ and $\lambda_\mu(\theta; 1)$ are the monotonic functions determined by \mathfrak{x}_0 and \mathfrak{x}_1 respectively:

$$(11) \quad \mathfrak{x}_\mu(\theta; 0) = g_\mu[\lambda_\mu(\theta; 0)], \quad \mathfrak{x}_\mu(\theta; 1) = g_\mu[\lambda_\mu(\theta; 1)].$$

The potential surfaces $\mathfrak{x}(t)$, $0 \leq t \leq 1$, form a "linear" path joining \mathfrak{x}_0 and \mathfrak{x}_1 .

LEMMA 3. For the linear path $\mathfrak{x}(t)$ constructed above,

$$\left| \frac{\mathfrak{x}(t') - \mathfrak{x}(t'')}{t' - t''} \right| \leq \frac{L}{2\pi} \cdot \text{maximum}_{\substack{0 \leq \theta \leq 2\pi \\ \mu=1, \dots, k}} |\lambda_\mu(\theta; 1) - \lambda_\mu(\theta; 0)|.$$

PROOF. The potential surface $\frac{\mathfrak{x}(t') - \mathfrak{x}(t'')}{t' - t''}$ has boundary values on C_μ equal to

$$\frac{g_\mu(\psi') - g_\mu(\psi'')}{t' - t''} = \frac{g_\mu(\psi') - g_\mu(\psi'')}{\psi' - \psi''} \cdot \frac{\psi' - \psi''}{t' - t''}$$

where $\psi' = (1 - t')\lambda_\mu(\theta; 0) + t'\lambda_\mu(\theta; 1)$, and ψ'' is a similar expression involving t'' . Using (9) one obtains

$$\left| \frac{g_\mu(\psi') - g_\mu(\psi'')}{t' - t''} \right| \leq \frac{L}{2\pi} \cdot |\lambda_\mu(\theta; 1) - \lambda_\mu(\theta; 0)|.$$

Similarly for the boundary values of $\frac{\mathfrak{x}(t') - \mathfrak{x}(t'')}{t' - t''}$ on any point circle c_μ . The lemma follows.

THEOREM 7. Let \mathfrak{x}_0^n and \mathfrak{x}_1^n , both having the domain G^n , $n = 1, 2, \dots$, be two sequences of potential surfaces in \mathfrak{B} converging to the same potential surface \mathfrak{x} . Let $\mathfrak{x}^n(t)$ be the linear path joining \mathfrak{x}_0^n and \mathfrak{x}_1^n . Then

$$\frac{E[\mathfrak{x}^n(t')] - E[\mathfrak{x}^n(t'')]}{t' - t''} \rightarrow 0 \quad \text{as } n \rightarrow \infty,$$

uniformly in t' and t'' .

PROOF. The desired difference quotient can be written in the form

$$(12) \quad \begin{aligned} \frac{E[\mathfrak{x}^n(t')] - E[\mathfrak{x}^n(t'')]}{t' - t''} &= \frac{E[\mathfrak{x}^n(t') - \mathfrak{x}^n(t''), \mathfrak{x}^n(t') + \mathfrak{x}^n(t'')]}{t' - t''} \\ &= E \left[\frac{\mathfrak{x}^n(t') - \mathfrak{x}^n(t'')}{t' - t''}, \mathfrak{x}^n(t') + \mathfrak{x}^n(t'') \right]. \end{aligned}$$

Now the monotonic functions $\lambda_\mu^n(\theta; 0)$ and $\lambda_\mu^n(\theta; 1)$ determined by \mathfrak{x}_0^n and \mathfrak{x}_1^n both converge uniformly to the monotonic function determined by \mathfrak{x} . Consequently $\mathfrak{x}^n(t)$ converges to \mathfrak{x} , uniformly in t ; and by Lemma 3 $\frac{\mathfrak{x}^n(t') - \mathfrak{x}^n(t'')}{t' - t''}$ converges to the potential surface identically equal to zero, uniformly in t' and t'' . Theorem 6 of Part I applied to the last expression in (12) shows that the desired difference quotient converges to $E[0, 2\mathfrak{x}] = 0$, uniformly in t' and t'' .

8. The Reducibility Condition

We shall obtain reducibility on the assumption that reducibility holds for the case of single boundaries. Such results for single boundaries have been derived in rather general cases by the author, and by Morse and Tompkins.

Suppose that the following result has been established for each of the curves Γ_μ , $\mu = 1, 2, \dots, k$.

Reducibility Condition for Single Boundaries.⁸ Let η be any surface in \mathfrak{P}^r . To each surface ξ of \mathfrak{P}^r there can be associated a linear path $\xi(t)$ joining ξ to $\xi' = \xi(1)$, where ξ' depends continuously on ξ , and $\xi' = \eta$ if $\xi = \eta$. For any η^0 and γ there is an α independent of γ and a δ such that, if ξ is in the δ -neighborhood of η , this linear path $\xi(t)$ has the following properties:

- 1). $D[\xi(t)]$ exists and depends continuously on t in $0 \leq t \leq 1$.
- 2). $D[\xi(t')] > D[\eta] + \eta$ and $D[\xi(t'')] > D[\eta] + \eta$ imply that

$$\frac{\Delta D}{\Delta t} = \frac{D[\xi(t')] - D[\xi(t'')]}{t' - t''} \leq -\alpha.$$

- 3). $\frac{\Delta D}{\Delta t} \leq \gamma$ for any t', t'' .

In case $\frac{dD[\xi(t)]}{dt}$ exists for $0 < t < 1$, Properties 2) and 3) are equivalent to:

- 2'). $D[\xi(t')] > D[\eta] + \eta$ implies that $\frac{dD[\xi(t')]}{dt} \leq -\alpha$.

- 3'). $\frac{dD[\xi(t)]}{dt} \leq \gamma$ for $0 < t < 1$.

There is an important consequence of Property 2).

LEMMA 4. If $D[\xi(t')] \leq D[\eta] + \eta$ for some t' , then $D[\xi(t)] \leq D[\eta] + \eta$ for all $t \geq t'$. If $D[\xi(t'')] > D[\eta] + \eta$ for some t'' , then $D[\xi(t)] > D[\eta] + \eta$ for all $t \leq t''$.

PROOF. Let the maximum of $D[\xi(t)]$ for $t \geq t'$ be attained for $t = \tau$. If $D[\xi(\tau)] > D[\eta] + \eta$, choose τ' between t' and τ and so near τ that $D[\xi(\tau')] > D[\eta] + \eta$. Then $\frac{D[\xi(\tau)] - D[\xi(\tau')]}{\tau - \tau'} \geq 0$, contrary to Property 2). Hence $D[\xi(\tau)] \leq D[\eta] + \eta$, and the first statement of the lemma is proved. The second statement of the lemma is a consequence of the first statement.

Therefore, the possibilities for the graph of $D[\xi(t)]$ as a function of t in $0 \leq t \leq 1$ are:

- a) If $D[\xi] \leq D[\eta] + \eta$, then $D[\xi(t)]$ is always at most $D[\eta] + \eta$.
- b) If $D[\xi] > D[\eta] + \eta$, then $D[\xi(t)]$ changes at a rate $\leq -\alpha$ until $D[\eta] + \eta$ is reached (if it is reached at all) and thereafter $D[\xi(t)]$ remains below $D[\eta] + \eta$. This is exactly what is meant by 'reducibility'. Compare page 36 of [6], Lemma 9 and Theorem 4 of [12], and page 445 of [7].

We now return to the case of several boundaries.

⁸ The reducibility condition stated here can be easily generalized but it is useless to do so.

⁹ All the quantities $\eta, \gamma, \alpha, \delta$ are positive.

THEOREM 8. *Suppose that the reducibility condition above has been established for the individual rectifiable curves $\Gamma_1, \Gamma_2, \dots, \Gamma_k$. Then the same condition holds for the case of the k boundaries $\Gamma_1, \Gamma_2, \dots, \Gamma_k$ together, i.e., in the space $\mathfrak{P}^{\Gamma_1, \Gamma_2, \dots, \Gamma_k}$.*

PROOF. Let η be a fixed surface in $\mathfrak{P} = \mathfrak{P}^{\Gamma_1, \Gamma_2, \dots, \Gamma_k}$ with boundary values η_μ , and let ξ with domain G and boundary values ξ_μ be any surface in \mathfrak{P} . To the potential surface ξ_μ with boundary Γ_μ there is associated the linear path $\xi_\mu(t)$ determined by our hypothesis for Γ_μ . The required linear path associated to ξ is defined over the domain G and has the boundary values $\xi_\mu(t), \mu = 1, \dots, k$. We have

$$(13) \quad D[\xi(t)] = \sum_{\mu=1}^k D_0[\xi_\mu(t)] + E[\xi(t)],$$

which shows that $D[\xi(t)]$ is a continuous function of t because of Property 1) of our hypothesis and of Theorem 5 in Part I.

Let η be arbitrary. Our hypothesis for Γ_μ determines a constant α_μ belonging to $\eta/2k$ in place of η . Set $\alpha = \text{minimum } \alpha_\mu \text{ for } \mu = 1, \dots, k$. Choose any $\gamma \leq \alpha/2k$. There is a δ_μ such that all the assertions of our hypothesis (with $\eta/2k$ replacing η) apply if ξ_μ is in the δ_μ -neighborhood of η_μ . Set $\delta' = \text{minimum } \delta_\mu \text{ for } \mu = 1, \dots, k$.

If ξ is close to η then $\xi(t)$ is likewise close to η (since $\xi_\mu(t)$ is close to η_μ for each μ). By Theorem 5 in Part I, there is a δ'' such that

$$(14) \quad |E[\xi(t)] - E[\eta]| < \eta/2$$

whenever ξ is in the δ'' -neighborhood of η . By Theorem 7, there is a δ''' such that

$$(15) \quad \left| \frac{\Delta E}{\Delta t} \right| = \left| \frac{E[\xi(t')] - E[\xi(t'')]}{t' - t''} \right| < \gamma \leq \alpha/2k$$

if ξ is in the δ''' -neighborhood of η .

The required δ is defined by $\delta = \text{minimum } \delta', \delta'', \delta'''$. Consider only surfaces ξ which are in the δ -neighborhood of η .

Suppose that $D[\xi(t')] > D[\eta] + \eta$ and $D[\xi(t'')] > D[\eta] + \eta$, where $t'' < t'$. The first inequality requires that $D_0[\xi_\nu(t')] > D_0[\eta_\nu] + \eta/2k$ for some $\mu = \nu$, by (13) and (14). Lemma 4 yields $D_0[\xi_\nu(t'')] > D[\eta_\nu] + \eta/2k$. Properties 2) and 3) of our hypotheses give

$$(16) \quad \frac{\Delta D_0[\xi_\nu]}{\Delta t} = \frac{D_0[\xi_\nu(t')] - D_0[\xi_\nu(t'')]}{t' - t''} \leq -\alpha,$$

and

$$(17) \quad \frac{\Delta D_0[\xi_\mu]}{\Delta t} \leq \gamma \leq \alpha/2k \quad \text{for all } \mu.$$

A final calculation, using (15), (16), (17), shows that

$$\begin{aligned}\frac{\Delta D[\mathbf{x}]}{\Delta t} &= \frac{D[\mathbf{x}(t')] - D[\mathbf{x}(t'')]}{t' - t''} = \sum_{\mu=1}^k \frac{\Delta D[\mathbf{x}_\mu]}{\Delta t} + \frac{\Delta E[\mathbf{x}]}{\Delta t} \\ &\leq -\alpha + \sum_{\mu \neq \nu} \frac{\alpha}{2k} + \frac{\alpha}{2k} = -\frac{\alpha}{2}.\end{aligned}$$

This proves Property 2).

Property 3) likewise holds, by (15) and (17). But this is of no importance in connection with reducibility. q.e.d.

For single boundaries reducibility has been proved by the author, by Morse and Tompkins, and by Courant for special classes of boundary curves. We shall discuss each briefly.

The author in [12] used the following type of path joining two boundary representations $\mathbf{x}_0(\theta)$ and $\mathbf{x}_1(\theta)$ of the curve Γ . Let $\mathbf{x}_0(\lambda_0(\theta)) = \mathbf{g}(\theta)$ and $\mathbf{x}_1(\lambda_1(\theta)) = \mathbf{g}(\theta)$ where $\lambda_0(\theta)$ and $\lambda_1(\theta)$ are monotonic functions of θ and $\mathbf{g}(\theta)$ is that representation of Γ the parameter θ of which is proportional to the arc length. Define $\mathbf{x}(\varphi; t)$ by

$$\mathbf{x}(\varphi; t) = \mathbf{g}(\theta) \quad \text{where} \quad \varphi = (1 - t)\lambda_0(\theta) + t\lambda_1(\theta).$$

Compare with the linear path (10), (11) used in the present paper. For this path a property such as Theorem 7 is very difficult to prove. Accordingly, the results of [12] cannot be used here.

Morse and Tompkins in [7] proved the reducibility condition stated above in case Γ has the following property:

$$(18) \quad \mathbf{g}'(\theta) \text{ exists and } |\mathbf{g}'(\theta_1) - \mathbf{g}'(\theta_2)| \leq M |\theta_1 - \theta_2|$$

for all θ_1, θ_2 . See Lemma 5.1, Lemma 5.4 and Theorem 5.1 in [7]. The hypothesis of Theorem 9 is consequently satisfied if each Γ_μ , $\mu = 1, \dots, k$, has this property.

The above reducibility condition is implicitly contained in the work of Courant [3] for polygonal boundaries. We shall not stress this point however, since Courant's method is directly applicable without the intervention of §§7, 8. We consider this matter in the next section.

9. The Case of Polygonal Boundaries

We shall show how to apply the method of Courant [3] to the case of several polygonal boundaries, using only the results of Part I.

Let the vertices of Γ_μ be $P_\mu, Q_\mu, R_\mu, S_\mu, T_\mu, \dots$ etc. For a given surface $\mathbf{x}(u, v)$ of \mathfrak{B} denote points on C_μ which are mapped into $P_\mu, Q_\mu, R_\mu, S_\mu, \dots$ by $p_\mu, q_\mu, r_\mu, s_\mu, \dots$. Map the circle C_μ by a linear transformation into a unit circle so that the images of p_μ, q_μ, r_μ are the specified points $\theta_1, \theta_2, \theta_3$. The images of s_μ, t_μ, \dots on the unit circle will be denoted collectively by σ_μ ; and $\sigma_1, \sigma_2, \dots, \sigma_k$ will be indicated collectively by σ .

Let Ω denote the space of elements $\{G, \sigma\}$, of domains G and points σ , with an obvious metric. The dimension of Ω is $3k - 6 + V$ where V is the total number of vertices in the polygons $\Gamma_1, \dots, \Gamma_k$. Ω has the same connectivity numbers as \mathfrak{R}_k .

The space of potential surfaces $\mathfrak{z}(u, v)$ of \mathfrak{P} defined over elements $\{G, \sigma\}$ of Ω will be indicated by \mathfrak{P}' . \mathfrak{P}' has the same connectivity numbers as \mathfrak{P} .

For a given $\{G, \sigma\}$ of Ω consider the problem of minimizing $D[\mathfrak{z}]$ among all surfaces \mathfrak{z} of \mathfrak{P}' defined over this $\{G, \sigma\}$. By the compactness of \mathfrak{P}_N (and of \mathfrak{P}'_N), this problem has a solution \mathfrak{z} . As in [3], using the linear path $\mathfrak{z}(t) = (1 - t)\mathfrak{z}_0 + t\mathfrak{z}_1$ joining two surfaces \mathfrak{z}_0 and \mathfrak{z}_1 having the same $\{G, \sigma\}$, the solution is unique. Denote it by $\mathfrak{z}(G, \sigma)$, and set $d(G, \sigma) = D[\mathfrak{z}(G, \sigma)]$.

THEOREM 9. *The surface $\mathfrak{z}(G, \sigma)$ and the quantity $d(G, \sigma)$ depend continuously on $\{G, \sigma\}$.*

PROOF. Let $\{G^n, \sigma^n\} \rightarrow \{G, \sigma\}$. Abbreviate $\mathfrak{z}(G, \sigma)$ by \mathfrak{z} , and let \mathfrak{z}_μ be the potential surface defined over a unit circle with the boundary values determined by \mathfrak{z} on C_μ . Vary the points σ_μ and the surface \mathfrak{z}_μ as in [3] so that the σ_μ move into σ_μ^n . The varied potential surface \mathfrak{z}_μ^n is such that

$$(19) \quad \mathfrak{z}_\mu^n \rightarrow \mathfrak{z}_\mu \quad \text{and} \quad D_0[\mathfrak{z}_\mu^n] \rightarrow D_0[\mathfrak{z}_\mu] \quad \text{as } n \rightarrow \infty.$$

Let \mathfrak{z}^n be the potential surface defined over $\{G^n, \sigma^n\}$ with boundary values on C_μ determined by \mathfrak{z}_μ^n , $\mu = 1, \dots, k$. We have

$$(20) \quad D[\mathfrak{z}^n] = \sum_{\mu=1}^k D_0[\mathfrak{z}_\mu^n] + E[\mathfrak{z}^n] \rightarrow \sum_{\mu=1}^k D_0[\mathfrak{z}_\mu] + E[\mathfrak{z}] = D[\mathfrak{z}]$$

by (19) and the continuity of the functional $E[\mathfrak{z}]$ (Theorem 5 of Part I). Because $D[\mathfrak{z}^n] \geq d(G^n, \sigma^n)$, (20) yields

$$(21) \quad d(G, \sigma) \geq \limsup_{n \rightarrow \infty} d(G^n, \sigma^n).$$

In particular the quantities $d(G^n, \sigma^n)$ are uniformly bounded.

On the other hand, by the compactness of \mathfrak{P}_N , a subsequence of the surfaces $\mathfrak{z}(G^n, \sigma^n)$ converges to a potential surface \mathfrak{z}^∞ defined over $\{G, \sigma\}$. The lower semicontinuity of the Dirichlet functional gives

$$(22) \quad d(G, \sigma) \leq D[\mathfrak{z}^\infty] \leq \liminf_{n \rightarrow \infty} d(G^n, \sigma^n).$$

The relations (21), (22) establish the continuity of $d(G, \sigma)$.

The equality sign holds throughout (22), so that $D[\mathfrak{z}^\infty] = d(G, \sigma)$ or $\mathfrak{z}^\infty = \mathfrak{z}(G, \sigma)$. Thus $\mathfrak{z}(G^n, \sigma^n) \rightarrow \mathfrak{z}(G, \sigma)$, and the theorem is proved.

This theorem shows first that the set of surfaces $\mathfrak{z}(G, \sigma)$ is topologically equivalent to the space Ω . The symbol Ω will henceforth designate the space of these surfaces $\mathfrak{z}(G, \sigma)$. The theorem then asserts that $D[\mathfrak{z}]$ is continuous over Ω .

So far as Morse theory is concerned, our discussion may be limited to the space Ω .

10. The Variational Condition

In this section, no restriction will be made on the contours $\Gamma_1, \dots, \Gamma_k$.

A surface \mathfrak{r} in \mathfrak{P} is a minimal surface if the analytic function $\varphi(w) = (x_u - ix_v)^2 = x_u^2 - x_v^2 - 2ix_u x_v$ is identically zero in each of the regions over which \mathfrak{r} is defined. The purpose of this section is to show, when η is not a minimal surface, that a neighborhood of η can be deformed so as to decrease the value of the Dirichlet functional. If η is an ordinary surface in \mathfrak{P} , i.e., not degenerate, then this result is an immediate consequence of variations performed by Courant. The difficulty arises when η is degenerate; for this case a more detailed investigation is necessary.

Let η with domain G , be a surface in \mathfrak{P} not a minimal surface. We may suppose without loss of generality that $-2\eta_u \eta_v$ is not identically zero in at least one of the regions G' which compose G .¹⁰ Let a be any value $\geq D[\eta]$. Let \mathfrak{r} , with domain H , indicate a surface in \mathfrak{P}_a near η , and let H' be that region of H corresponding to G' and normalized analogously to G' . The following lemma is a result of performing certain variations due to Courant (see [2]):

LEMMA 5. Let \mathfrak{r} , with domain H , be a surface in \mathfrak{P}_a , and suppose that $-2\mathfrak{r}_u \mathfrak{r}_v \leq -b < 0$ in a fixed square K (with sides parallel to the u and v axes) of side 2τ interior to the region H' of H . Then a deformation $\mathfrak{r}(\epsilon)$, $0 \leq \epsilon \leq 1$, of \mathfrak{r} can be obtained such that

$$D[\mathfrak{r}(\epsilon)] \leq D[\mathfrak{r}] - \beta\epsilon,$$

where β depends on a , b , and τ . This deformation $\mathfrak{r}(\epsilon)$ is a continuous function of ϵ , and of \mathfrak{r} as long as no boundaries of H' appear in the square K .

PROOF. Deform the region H' into $H'(\epsilon)$ by the transformation

$$(23) \quad \begin{cases} U = u + \epsilon\lambda(u, v), \\ V = v, \end{cases}$$

where

$$(24) \quad \begin{cases} \lambda(u, v) = 0 & \text{outside the square } K, \\ \lambda(u, v) = \frac{[\tau^2 - (u - u_0)^2][2\tau - (v - v_0)]}{2\tau} & \text{inside } K. \end{cases}$$

Here, (u_0, v_0) is the center of the lower side L of the square K and $0 \leq \epsilon \leq 1/(4\tau)$. The function $\lambda(u, v)$ is non-negative, is discontinuous across the side L , and has its first derivatives bounded in absolute value by 2τ . By coordinatizing the point u of the lower edge of L with the point $u + \epsilon\lambda(u, v_0)$ of the upper edge of L , the transformed domain becomes a Riemann domain. Define $\mathfrak{r}(\epsilon)$ over the Riemann domain $H'(\epsilon)$ as the potential surface having the same boundary

¹⁰ If $-2\mathfrak{r}_u \mathfrak{r}_v = 0$ but $\mathfrak{r}_u^2 - \mathfrak{r}_v^2 \neq 0$, rotate the (u, v) coordinate system 45° and consider the new coordinates u', v' . We have $u' = (u + v)/\sqrt{2}$, $v' = (u - v)/\sqrt{2}$ and $x_u^2 - x_v^2 = 2 x_{u'} x_{v'}$.

values as \mathfrak{z} on each boundary. One may map $H'(\epsilon)$ conformally on a circular domain normalized analogously to H' , and consider $\mathfrak{z}(\epsilon)$ as a potential surface over this circular domain.¹¹ In all the other regions besides H' which compose H , set $\mathfrak{z}(\epsilon) \equiv \mathfrak{z}$. It can be proved by use of conformal mapping methods that $\mathfrak{z}(\epsilon)$ depends continuously on ϵ , and on \mathfrak{z} as long as no boundaries of H' appear inside the square K .

To compute $D[\mathfrak{z}(\epsilon)]$, introduce the surface $\mathfrak{z}(\epsilon)$ defined over the region $H'(\epsilon)$ by $\mathfrak{z}(U, V; \epsilon) = \mathfrak{z}(u, v)$ where U, V are given in (23). Then $\mathfrak{z}(\epsilon)$ has the same boundary values as $\mathfrak{z}(\epsilon)$ in $H'(\epsilon)$, and we have by a simple calculation

$$D[\mathfrak{z}(\epsilon)] \leq D[\mathfrak{z}(\epsilon)] = D[\mathfrak{z}] + \frac{\epsilon}{2} \int_{u_0-\tau}^{u_0+\tau} \lambda(u, v_0)(-2\mathfrak{z}_u \mathfrak{z}_v)_{v=v_0} du + \epsilon^2 I$$

where $|I| \leq 4\tau^2 D[\mathfrak{z}] \leq 4\tau^2 a$. Since $-2\mathfrak{z}_u \mathfrak{z}_v \leq -b$ in the square K ,

$$\int_{u_0-\tau}^{u_0+\tau} \lambda(u, v_0)(-2\mathfrak{z}_u \mathfrak{z}_v)_{v=v_0} du \leq -b \int_{u_0-\tau}^{u_0+\tau} [\tau^2 - (u - u_0)^2] du = -\frac{4\tau^3 b}{3}.$$

For $0 \leq \epsilon \leq \sigma'$, where σ' is the smaller of the two numbers $1/(4\tau)$, $\tau b/(12a)$ we obtain

$$\frac{1}{2} \int_{u_0-\tau}^{u_0+\tau} \lambda(u, v_0)(-2\mathfrak{z}_u \mathfrak{z}_v)_{v=v_0} du + \epsilon I \leq -\frac{4\tau^3 b}{3} + \frac{\tau b}{12a} \cdot 4\tau^2 a = -\tau^3 b.$$

Hence, in $0 \leq \epsilon \leq \sigma'$, $D[\mathfrak{z}(\epsilon)] \leq D[\mathfrak{z}] - \tau^3 b \epsilon$. Replacing ϵ by $\sigma' \epsilon$, the lemma is obtained with $\beta = \tau^3 b \sigma'$.

Lemma 5 will now be used to perform a piecewise deformation of a neighborhood of \mathfrak{y} in \mathfrak{B}_a . There is an open region in G' in which $-2\eta_u \eta_v \leq -2b < 0$. If \mathfrak{y} is ordinary, a square K can be selected in this open region, and Lemma 5 immediately yields the required deformation. If \mathfrak{y} is degenerate, Lemma 5 does not apply since a surface near \mathfrak{y} may have small circular boundaries in K ; but it must have less than k small boundaries. Accordingly, in the region where $-2\eta_u \eta_v \leq -2b$, select k squares K_1, K_2, \dots, K_k of sufficiently small side 2τ with sides parallel to the u and v axes, and at a distance $\geq 16\tau$ from each other and from the boundaries of G' . Consider the 2δ -neighborhood of \mathfrak{y} such that, for any \mathfrak{z} (with domain H) in this neighborhood:

1) those boundaries of H' which correspond to the boundaries of G' are displaced from them by at most a distance τ , and the other boundaries of H' (these will hereafter be called *small* boundaries) have a radius $\leq \tau$;

2) if there are no boundary points of H' within a 2τ neighborhood of the center of the square K_j , then $-2\mathfrak{z}_u \mathfrak{z}_v \leq -b$ throughout this K_j . The desired

¹¹ Compare [2]. The end points of L are actually transformed into points (and not merely slits). Of course it is possible to obtain a similar lemma and the variational condition by using variations not depending on the theory of conformal mapping, as in [1], [2]. But it would then be necessary to distinguish two cases: varying boundary values and varying the domain.

neighborhood of η in \mathfrak{P}_a which will be deformed is the closed δ -neighborhood $M_1 = N_\delta$ in \mathfrak{P}_a .

The piecewise deformation of M_1 will be obtained by applying Lemma 5 successively for each square K_j . For this purpose, limit ϵ to the range $0 \leq \epsilon \leq \sigma$ where σ is so small that:

3) the surfaces $\mathfrak{r}(\epsilon)$ of Lemma 5 for each square K_j is displaced from \mathfrak{r} at most a distance δ/k ;

4) the boundaries of $H'(\epsilon)$ are displaced from the corresponding boundaries of H' at most a distance τ/k .

It follows from these conditions that any surface \mathfrak{z} obtained by a k -fold repeated application of all these deformations, satisfies 3), 4) with δ, τ replacing $\delta/k, \tau/k$. Because of this, conditions 1), 2) likewise apply to \mathfrak{z} . Replace ϵ by $\sigma\epsilon$, and $\sigma\beta$ by β so that Lemma 5 and the above apply for $0 \leq \epsilon \leq 1$.

Let A_1 be that subset of M_1 consisting of those surfaces \mathfrak{r} interior to M_1 whose region H' contains boundaries at a distance $< 3\tau$ from the center of K_1 . Indicate the closure of A_1 by \bar{A}_1 , and set $B_1 = M_1 - \bar{A}_1$. If \mathfrak{r} is any surface in A_1 , it remains fixed in the deformation; if \mathfrak{r} is in B_1 , construct the deformation $\mathfrak{r}(\epsilon)$ of Lemma 5 for the square K_1 . Indicate the boundary operator by \mathfrak{B}_1 (boundary as subset of \mathfrak{P}_a), the deformation operator by \mathfrak{D}_1 , and the final image $\mathfrak{r}(1)$ by \mathfrak{F}_1 . Then

$$(25) \quad \begin{cases} \mathfrak{B}\mathfrak{D}_1 \bar{B}_1 = \bar{B}_1 - \mathfrak{D}_1 \mathfrak{B}_1 \bar{B}_1 - \mathfrak{F}_1 \bar{B}_1, \\ \text{or } \bar{B}_1 \sim \mathfrak{D}_1 \mathfrak{B}_1 \bar{B}_1 + \mathfrak{F}_1 \bar{B}_1. \end{cases}$$

These relations and the operators $\mathfrak{B}_1, \mathfrak{D}_1, \mathfrak{F}_1$ are understood in the sense that they apply in a natural way to any Vietoris chain on \bar{B}_1 . In view of Lemma 5, all the surfaces involved in (25) lie on \mathfrak{P}_a , and $\mathfrak{F}_1 \bar{B}_1$ lies on $\mathfrak{P}_{a-\beta}$; the homology¹² and all subsequent homologies, take place over \mathfrak{P}_a . It follows from (25) that¹²

$$(26) \quad \begin{cases} M_1 \sim \bar{A}_1 + \mathfrak{D}_1 \mathfrak{B}_1 \bar{B}_1 + \mathfrak{F}_1 \bar{B}_1 = M_2, \\ \mathfrak{P}_a \sim (\overline{\mathfrak{P}_a - M_1}) + M_2 = P_2. \end{cases}$$

The surfaces in M_2 may be designated by $f(\mathfrak{r}, t)$ in place of $\mathfrak{r}(\epsilon)$, where $0 \leq t \leq 1$ if \mathfrak{r} belongs to $\mathfrak{B}_1 \bar{B}_1$, $t = 0$ if \mathfrak{r} belongs to A_1 , and $t = 1$ if \mathfrak{r} is in the interior of B_1 . The surfaces in P_2 consist of M_2 and $f(\mathfrak{r}, 0)$ where \mathfrak{r} belongs to $\mathfrak{P}_a - M_1$. The surface \mathfrak{r} , from which $f(\mathfrak{r}, t)$ was obtained, is called the *pre-image* of the surface $f(\mathfrak{r}, t)$. The set P_2 may be considered as a new space in which distance between $f(\mathfrak{r}_1, t_1)$ and $f(\mathfrak{r}_2, t_2)$ is defined by $|\mathfrak{r}_1 - \mathfrak{r}_2| + |t_1 - t_2|$, where $|\mathfrak{r}_1 - \mathfrak{r}_2|$ is the distance between $\mathfrak{r}_1, \mathfrak{r}_2$ in \mathfrak{P}_a .

Suppose that M_j and P_j ($j \leq k$) have been constructed, and consist of surfaces in \mathfrak{P}_a designated by $f(\mathfrak{r}, t_1, \dots, t_{j-1})$ where $0 \leq t_\mu \leq 1$ or $t_\mu = 0$ or $t_\mu = 1$ according to the situation of \mathfrak{r} in certain subsets of \mathfrak{P}_a . Let A_j be the subset

¹² Any Vietoris chain in M_1 is homologous by subdivision to a Vietoris chain part of which lies wholly in \bar{A}_1 and the other part wholly in \bar{B}_1 . The homology $M_1 \sim M_2$ is a consequence. Similarly for $\mathfrak{P}_a \sim P_2$.

of M_j consisting of all the surfaces $f(\mathfrak{x}, t_1, \dots, t_{j-1})$ in M_j where the region H' for \mathfrak{x} contains boundaries at a distance $< 3\tau$ from the center of the square K_j ; set $B_j = M_j - \bar{A}_j$. If $f(\mathfrak{x}, t_1, \dots, t_{j-1})$ is any surface in A_j , it remains fixed. If $f(\mathfrak{x}, t_1, \dots, t_{j-1})$ is in B_j , deform $f(\mathfrak{x}, t_1, \dots, t_{j-1})$, considered as a surface in \mathfrak{P}_a , according to Lemma 5 for the square K_j ; Lemma 5 applies in view of conditions 3), 4) and 1), 2) above. Indicate the deformed surface by $f(\mathfrak{x}, t_1, \dots, t_{j-1}, t_j)$ where $0 \leq t_j \leq 1$. As in (25), (26),

$$(27) \quad \begin{cases} M_j \cup \bar{A}_j + \mathfrak{D}_j \mathfrak{B}_j \bar{B}_j + \mathfrak{F}_j \bar{B}_j = M_{j+1}, \\ P_j \cup \overline{(\mathfrak{P}_a - M_1)} + M_{j+1} = P_{j+1}, \end{cases}$$

where closure \bar{A}_j or \bar{B}_j , boundary \mathfrak{B}_j and image \mathfrak{F}_j are taken as subsets of P_j . All the relations (27) for $j = 1, 2, \dots, k$ yield

$$(28) \quad \begin{cases} M_1 \cup M_{k+1}, \\ \mathfrak{P}_a \cup P_{k+1}. \end{cases}$$

It is clear from Lemma 5 that if $\mathfrak{z} = f(\mathfrak{x}, t_1, t_2, \dots, t_k)$ is any surface in M_{k+1} , then

$$(29) \quad D[\mathfrak{z}] \leq D[\mathfrak{x}] - \beta \cdot \sum_{\mu=1}^k t_\mu.$$

Furthermore, we have

LEMMA 6. *Let \mathfrak{x} be any surface in the interior of M_1 , and $\mathfrak{z} = f(\mathfrak{x}, t_1, t_2, \dots, t_k)$ any surface in M_{k+1} with \mathfrak{x} as pre-image. Then $t_\gamma = 1$ for at least one t_γ of t_1, t_2, \dots, t_k .*

PROOF. If the lemma were false, there would exist a $\mathfrak{z} = f(\mathfrak{x}, t_1, t_2, \dots, t_k)$ in M_{k+1} for which \mathfrak{x} is interior to M_1 and $t_\mu \neq 1$, $\mu = 1, 2, \dots, k$. Since $t_k \neq 1$, it follows that the surface $f(\mathfrak{x}, t_1, \dots, t_{k-1})$ belongs to \bar{A}_k . Hence there are surfaces $f(\mathfrak{x}', t'_1, \dots, t'_{k-1})$ belonging to A_k which are arbitrarily near $f(\mathfrak{x}, t_1, \dots, t_{k-1})$; \mathfrak{x}' , being near \mathfrak{x} , is still interior to M_1 and t'_1, \dots, t'_{k-1} are $\neq 1$. In particular, arbitrarily near the region H' for \mathfrak{x} are regions H'_1 for \mathfrak{x}' which contain boundaries at a distance $< 3\tau$ from the center of K_k .

Since $t'_{k-1} \neq 1$, the surface $f(\mathfrak{x}', t'_1, \dots, t'_{k-2})$ belongs to \bar{A}_{k-1} . As previously, there are surfaces $f(\mathfrak{x}'', t''_1, \dots, t''_{k-2})$ arbitrarily near $f(\mathfrak{x}', t'_1, \dots, t'_{k-2})$ which belong to A_{k-1} . In particular, arbitrarily near H'_1 are regions H'_2 (for \mathfrak{x}'') which contain boundaries at a distance $< 3\tau$ from the center of K_{k-1} . Because H'_1 has a similar property for the square K_k , it follows that H'_2 contains boundaries at a distance $< 3\tau$ from the centers of K_{k-1} and of K_k . Continuing in this way, one finally obtains regions H'_k (which are regions for surfaces interior to M_1) with boundaries at a distance $< 3\tau$ from each of the centers of K_1, K_2, \dots, K_k . But this contradicts the facts that H'_k contains less than k small circles, each

small circle has a radius $\leq \tau$ (see 1), 2) above), and the squares K_1, K_2, \dots, K_k have a distance $\geq 16\tau$ from each other. The lemma is established.

The results, (28), (29) and lemma 6, of this piecewise deformation are summarized in

THEOREM 10. *Let η be a surface in \mathfrak{P} not a minimal surface, and a any value $\geq D[\eta]$. There are two positive constants δ and β , and a piecewise deformation of \mathfrak{P}_a in itself which yields*

$$\mathfrak{P}_a \rightsquigarrow P$$

(the homology taking place in \mathfrak{P}_a and applying to any Vietoris cycle on \mathfrak{P}_a). Here, P is a subspace of \mathfrak{P}_a consisting of surfaces z of the form $z = f(x, t_1, t_2, \dots, t_k)$, where t_μ , $\mu = 1, 2, \dots, k$, takes the value 0, or values between 0 and 1, or the value 1 according to the situation of x in \mathfrak{P}_a , and $f(x, t_1, \dots, t_k)$ has the following properties:

- 1) $f(x, t_1, \dots, t_k)$ is continuous in x, t_1, \dots, t_k ; and $f(x, 0, 0, \dots, 0) = x$.
- 2) If $|x - \eta| > \delta$, then t_1, \dots, t_k take only the value 0.
- 3) If $|x - \eta| = \delta$, then $D[z] \leq D[x] - \beta \cdot \sum_{\mu=1}^k t_\mu$.
- 4) If $|x - \eta| < \delta$, then $D[z] \leq D[x] - \beta$.

The above theorem applies automatically to \mathfrak{P}' and Ω as well as \mathfrak{P} . Throughout theorem 10 replace \mathfrak{P} and P by Ω and Q .

11. The Main Theorems. Remarks

On the basis of the reducibility condition (Theorem 8 or Theorem 9) and the variational condition (Theorem 10), it is easy to prove that on each k -cap with cap limit a there is a *minimal* surface x for which $D[x] = a$. For the meaning of these terms, and such a proof, see [6] and Theorem 6 of [12]. This establishes the validity of the Morse theory.

MAIN THEOREM I. *Let $\Gamma_1, \Gamma_2, \dots, \Gamma_k$ be k non-intersecting closed rectifiable Jordan curves in space. Suppose that the reducibility condition of §8, page 213 has been established for each curve Γ_μ individually. Then the Morse theory applies to the minimal surfaces (degenerate as well as ordinary) bounded by $\Gamma_1, \dots, \Gamma_k$.*

There remains the question of determining the connectivity numbers of \mathfrak{P} , where only those Vietoris cycles are considered which lie on \mathfrak{P}_N for sufficiently large N . This is reduced by Theorems 2, 3 of Part I to the corresponding question for \mathfrak{P}^Γ . The connectivity numbers of \mathfrak{P}^Γ have been determined for rather general classes of curves Γ in [12], [7], [8]. They are: $R_0 = 1, R_1 = R_2 = \dots = R_n = \dots = 0$. Hence in these cases the connectivity numbers of \mathfrak{P} are likewise $R_0 = 1, R_1 = R_2 = \dots = R_n = \dots = 0$, by theorems 2, 3.

MAIN THEOREM II. *Let $\Gamma_1, \Gamma_2, \dots, \Gamma_k$ be k non-intersecting simple closed polygons in space. Then the Morse theory applies to the minimal surfaces (degenerate as well as ordinary) bounded by $\Gamma_1, \dots, \Gamma_k$. In particular, if M_n is*

the sum of the n^{th} type numbers of all blocs of minimal surfaces bounded by $\Gamma_1, \Gamma_2, \dots, \Gamma_k$, and if each M_n is finite, then

$$(30) \quad \left\{ \begin{array}{l} M_0 \geq 1, \\ M_1 - M_0 \geq -1, \\ \vdots \\ M_n - M_{n-1} + \dots + (-1)^n M_0 \geq (-1)^n, \\ \vdots \end{array} \right.$$

Also, $M_n = 0$ for all $n > 3k - 6 + V$, where V is the total number of vertices in all the polygons Γ_μ , $\mu = 1, \dots, k$.

Main Theorem II is a result of Part I, and §§9, 10. No use of §§7, 8 need be made.

An application of the Main Theorem I is to the case discussed by Morse and Tompkins in [7]. If each Γ_μ has the property (18), page 215, the Morse theory and the inequalities (30) apply.

Since degenerate as well as ordinary minimal surfaces are included in the main theorems above there are many more possibilities than in the case of one boundary. This leads to problems of the following kind. Suppose that a degenerate minimal surface \mathfrak{r} consists of two pieces, one \mathfrak{r}' bounded by a set (Γ') of the boundaries and the other \mathfrak{r}'' bounded by the remaining set (Γ'') . What is the relation, in the form of inequalities, between the Morse type of \mathfrak{r} in the space $\mathfrak{P}^{(\Gamma)}$ and the Morse types of \mathfrak{r}' in $\mathfrak{P}^{(\Gamma')}$ and of \mathfrak{r}'' in $\mathfrak{P}^{(\Gamma'')}$?

COLLEGE OF THE CITY OF NEW YORK

BIBLIOGRAPHY

1. COURANT, R., *Plateau's problem and Dirichlet's principle*, Annals of Math., 38 (1937), pp. 679-724.
2. COURANT, R., *The existence of minimal surfaces of given topological structure under prescribed boundary conditions*, Acta Math., 72 (1940), pp. 51-98.
3. COURANT, R., *Critical points and unstable minimal surfaces*, Proc. Nat. Acad. Sci., 27 (1941), pp. 51-57.
4. DOUGLAS, J., *Solution of the problem of Plateau*, Trans. Amer. Math. Soc., 33 (1931), pp. 263-321.
5. DOUGLAS, J., *Minimal surfaces of higher topological structure*, Annals of Math., 40 (1939), pp. 205-298.
6. MORSE, M., *Functional topology and abstract variational theory*, Mémorial des Sci. Math., vol. 92 (1939).
7. MORSE, M., AND TOMPKINS, C., *The existence of minimal surfaces of general critical type*, Annals of Math., 40 (1939), pp. 443-472.
8. MORSE, M., AND TOMPKINS, C., *Minimal surfaces not of minimum type by a new mode of approximation*, Annals of Math., 42 (1941), pp. 62-72.
9. MORSE, M., AND TOMPKINS, C., *Unstable minimal surfaces of higher topological structure*, Duke Math. Journ., 8 (1941), pp. 350-375.
10. RADÓ, T., *On the problem of Plateau*, Ergeb. der Math., vol. 2 (1933).
11. SHIFFMAN, M., Bull. Amer. Math. Soc., 44 (1938), p. 637, abstract of a paper read to the society in Sept., 1938.
12. SHIFFMAN, M., *The problem of Plateau for non-relative minima*, Annals of Math., 40 (1939), pp. 834-854.

ON THEORIES WITH A COMBINATORIAL DEFINITION OF "EQUIVALENCE"

By M. H. A. NEWMAN

(Received June 23, 1941)

1

The name "combinatorial theory" is often given to branches of mathematics in which the central concept is an equivalence relation defined by means of certain "allowed transformations" or "moves." A class of objects is given, and it is declared of certain pairs of them that one is obtained from the other by a "move"; and two objects are regarded as "equivalent" if, and only if, one is obtainable from the other by a series of moves. For example, in the theory of free groups the objects are words made from an alphabet $a, b, \dots, a^{-1}, b^{-1}, \dots$, and a move is the insertion or removal of a consecutive pair of letters xx^{-1} or $x^{-1}x$. In combinatorial topology the objects are complexes, and the allowed moves are "breaking an edge" by the insertion of a new vertex, or the reverse of this process.¹ In Church's "conversion calculus"² the rules II and III are "moves" of this kind.

In many such theories the moves fall naturally into two classes, which may be called "positive" and "negative." Thus in the free group the cancelling of a pair of letters may be called a positive move, the insertion negative; in topology the breaking of an edge, in the conversion calculus the application of Rule II (elimination of a λ), may be taken as the positive moves. In theories that have this dichotomy it is always important to discover whether there is what may be called a "theorem of confluence," namely, whether if A and B are "equivalent" it follows that there exists a third object, C , derivable both from A and from B by positive moves only. A closely connected problem is the search for "end-forms," or "normal forms," i.e. objects which admit no positive move. It is obvious that in a theory in which the confluence theorem holds no equivalence class can contain more than one end-form, but there remains the question whether in such a class any random series of positive moves must terminate at the end-form, or whether infinite series of moves may also exist.

The purpose of this paper is to make a start on a general theory of "sets of moves" by obtaining some conditions under which the answers to both the above questions are favorable. The results are essentially about "partially-ordered" systems, i.e. sets in which there is a transitive relation $>$, and sufficient conditions are given for every two elements to have a lower bound (i.e. for the set to be "directed") if it is known that every two "sufficiently near" elements have a lower bound. What further conditions are required for the existence of a *greatest* lower bound is not relevant to the present purpose, and is reserved for a later discussion.

¹ See Alexander [1] and Newman [1].

² See Church [1] and references there given.

As an application the normal form theorem of Church and Rosser [1] in the conversion calculus is derived.

2

We are concerned with two kinds of entities, "objects" and the "moves" performed on them, and each move is associated with two objects, "initial" and "final." We are therefore dealing essentially with *indexed 1-complexes* (in which, therefore, a positive sense is assigned in each 1-cell), the vertices being the "objects," and the positive 1-cells the "moves." It will be convenient to make use of this topological terminology.³ The incidence relations are in no way restricted: there may be many cells with the same vertices, and the initial and final vertices of a cell may coincide. In diagrams the positive 1-cells slope down the page, and some of the terms used are chosen accordingly.

Vertices are denoted by italic letters, cells (the single word is used from now on for "positive 1-cell") by the letters ξ, η, ζ, ω with various suffixes. " $x\mu y$ " means "there is a cell with initial vertex x and final vertex y ." An ordered set of cells $\xi_1, \xi_2, \dots, \xi_k$, form a *path* π if there are vertices x_0, x_1, \dots, x_k such that x_{i-1} and x_i are the vertices of ξ_i for $1 \leq i \leq k$. The cell ξ_i is *direct* or *reversed* in π according as it runs from x_{i-1} to x_i or from x_i to x_{i-1} , and the path is denoted by $e_1\xi_1 + e_2\xi_2 + \dots + e_k\xi_k$, where e_i is ± 1 as ξ_i is direct or reversed. If there are no reversed cells, π is a *descending* path. It is convenient to regard a single vertex, x , as a "null path" with x as initial and final vertex. A vertex which is not the initial vertex of any cell is a *minimal vertex*, or *end*.

If there is at least one non-null descending path from x to y we write $x > y$. z is a *lower* (*upper*) *bound* of x and y if $x \geq z$ and $y \geq z$ (if $z \geq x$ and $z \geq y$).

3

Expressed in this terminology the confluence property is

(A) If x_1 and x_2 are connected by a path in the indexed complex Σ they have a lower bound.

By a simple induction on the number of cells in a path from x_1 to x_2 this property can be deduced from the following special case of it:

(B) If x_1 and x_2 have an upper bound they have also a lower bound.

This in its turn is easily deduced from the still more special form (C):

(C) If $a\mu x_1$ and $a > x_2$, x_1 and x_2 have a lower bound.

The transition from (B) to (C) is a step towards localizing the property, and the theorems that will be proved in this paper give conditions in which the localization may be completed, i.e. in which (A) may be inferred from the following condition (holding for all a , x_1 and x_2):

(D) If $a\mu x_1$ and $a\mu x_2$, x_1 and x_2 have a lower bound.

NOTE. The cell and vertex terminology, although the most convenient for

³ The notions that arise are closely related to those of the theory of partially ordered sets, but usually not identical. Except in the case of identity the terms of that theory are therefore avoided.

our purpose, may suggest that " $x\mu y$ " implies that y is a "next" vertex below x . Actually the force of μ is that y is an element satisfying $y < x$, and lying in a certain neighborhood of x . For example, all the conditions (A) to (D) are satisfied if the vertices are taken to be the points of a vertical plane, and the positive 1-cells the downward sloping directed segments of length less than 1.

4

The simplest way of strengthening (D) so that it implies (A), is to require that paths descending from x_1 and x_2 to their lower bound shall each contain one cell; or, in terms of moves, that if two moves are possible on an object X , they can also be performed one after the other, and give the same result in either order.

THEOREM 1. *Let Σ be such that if $a\mu x$ and $a\mu y$, and $x \neq y$, there exists b such that $x\mu b$ and $y\mu b$. Then property (A) holds.*

Let " $x\nu y$ " denote " $x\mu y$ or $x = y$." We prove that if $a\mu x$ and $a\mu y_1\mu y_2\mu \cdots \mu y_k$, there is a b_k such that $x\nu b_1\nu b_2\nu \cdots \nu b_k$ and $y_k\nu b_k$,—a stronger form of (C). Suppose this proved for $k - 1$ (the case $k = 1$ following immediately from the datum), and let $x\nu b_1\nu b_2\nu \cdots \nu b_{k-1}$, and $y_{k-1}\nu b_{k-1}$. If $y_{k-1} = b_{k-1}$ take b_k to be y_k . If $y_{k-1}\mu b_{k-1}$, since also $y_{k-1}\mu y_k$ there exists a b_k such that $b_{k-1}\nu b_k$ and $y_k\nu b_k$; and this completes the induction.

COROLLARY 1.1. *The theorem remains true if " $x\nu b$ and $y\nu b$ " is substituted for " $x\mu b$ and $y\mu b$ " in the enunciation. (No change is needed in the proof.)*

This almost trivial result is sufficient to settle many of the more familiar theorems of the kind that we are considering. In the "word groups" already referred to, a move is to be regarded as completely determined by the initial and final words, (so that e.g. $xx^{-1}x \rightarrow x$ is regarded as the same move whether the first or last two letters are cancelled). Hence two pairs $\mathbf{x}\mathbf{x}^{-1}$ and $\mathbf{y}\mathbf{y}^{-1}$ (where \mathbf{x} and \mathbf{y} may be of the form u^{-1}) in the same word \mathbf{W} , that give rise to different possible moves on \mathbf{W} , have no common letter and give the same result if cancelled in either order. Since every series of positive moves (cancellations) terminates it follows that all such series starting from a given word \mathbf{W} lead to a common end-form.

Theorems of the Jordan-Hölder type also belong to this category. The kernel of these theorems is a theorem on modular lattices (say with the partial ordering $>$ and the operations \vee and \wedge). If X, Y, Z are consecutive elements in a descending chain, \mathcal{S} , in such a lattice let the chain \mathcal{S}' obtained by substituting Y' for Y be said to be *directly related* to \mathcal{S} (\mathcal{S}' dr \mathcal{S}) if $X = X \vee Y'$ and $Z = Y \wedge Y'$; and \mathcal{S}' shall be *related* to \mathcal{S} if it is obtainable from \mathcal{S} by a succession of such steps. The theorem in question is then that *from any two finite descending chains, \mathcal{S} and \mathcal{S}' , from A to B , a pair of related chains \mathcal{S}_1 and \mathcal{S}'_1 , can be obtained by the insertion of a finite number of additional terms in \mathcal{S} and \mathcal{S}' respectively.* This is evidently a "confluence" theorem. To apply 1 we take as a typical vertex of Σ the class $[\mathcal{S}]$ of all chains related to a chain \mathcal{S} , and as a positive 1-cell the ordered pairs of classes $[\mathcal{S}_1], [\mathcal{S}_2]$, where \mathcal{S}_2 is obtained from \mathcal{S}_1

by the insertion of one additional term,—say P —between X and Y . Then if \mathcal{S}'_1 dr \mathcal{S}_1 , the insertion of a suitable term in \mathcal{S}_2 gives a chain \mathcal{S}'_2 related to \mathcal{S}'_1 ; and hence more generally any member of $[\mathcal{S}_1]$ can be made into a chain related to \mathcal{S}_2 , by the insertion of one suitable term. Two successive “positive moves” on $[\mathcal{S}_1]$ can therefore be represented by two successive insertions of new elements in the same chain \mathcal{S} , and evidently the order in which they are inserted does not affect the result. The system therefore fulfils the conditions of Theorem 1. But any two chains descending from A to B have an “upper bound” in Σ , namely the class $[AB]$. Therefore they have a “lower bound,” and this is the required result.

5

In these examples it is obvious that if an end-form exists it is reached by random descent. This is necessarily so in all systems with non-interference of moves:

THEOREM 2. *Under the conditions of Theorem 1, if there is a descending path of k cells from a to an end e , no descending path from a contains more than k cells.*

If $k = 1$, Σ cannot contain a cell ay with $y \neq e$, since if it does b exists such that $y\mu b$ and $e\mu b$, and e is not an end. In the general case let π be a descending path $\xi_1 + \xi_2 + \dots + \xi_k$ joining a to e , and let $\eta_1 + \eta_2 + \dots + \eta_j$ be any descending path from a . Let ξ_1 and η_1 be cells ax and ay . If $x = y$ it follows immediately from an induction that $j \leq k$. If not, let the cells ζ and ω descend from x and y to the common vertex w . By Theorem 1 there is a descending path σ from w to a vertex $\leq e$, i.e., since e is an end, to e itself. Since $\xi_2 + \dots + \xi_k$ has $k - 1$ cells, $\zeta + \sigma$ has, by an inductive hypothesis, at most $k - 1$ cells; therefore $\omega + \sigma$, and finally also $\eta_2 + \dots + \eta_j$, have at most $k - 1$ cells,—i.e. $j \leq k$.

COROLLARY 2.1. *Every descending path from a is part of a descending path of k cells from a to e (i.e. there is “random descent” to e).*

That Theorem 2 and Corollary 2.1 fail if the condition is weakened as in Corollary 1.1 is shown by the example in Fig. 1, (positive cells slope downward).

The main criteria for “confluence” are established in Theorems 3, 4, 5, and 9, all of which are independent. It is Theorems 5 and 9 that are used in the application to the conversion calculus.

THEOREM 3. *In an indexed complex in which all descending paths are finite, (D) implies (A).*

(Note that in such a complex “ $>$ ” is a proper ordering, since if $x > x$ an infinite descending path is obtained by going round and round the re-entrant path from x to x .)

⁴ Namely, if X and Y are in \mathcal{S}'_1 , insert P itself; if XYZ and $XY'Z$ are consecutive terms of \mathcal{S}_1 and \mathcal{S}'_1 respectively, insert $P' = Y' \wedge P$ in \mathcal{S}'_1 ; if UXY and $UX'Y$, insert $P'' = X' \vee P$. It is easily shewn that in the second case $XPYZ$ is related to $XY'P'Z$, in the third $UXPY$ to $UP''X'Y$. Cf. Birkhoff [1] p. 37.

The symbol $[\xi]_k$ is used as an abbreviation for $\xi_1 + \xi_2 + \cdots + \xi_k$. It is convenient to allow the value $k = 0$, $[\xi]_0$ being a null path.

A *peak* of a path is the common vertex of a successive pair of cells $-\xi + \eta$, ("up" before "down").

Let $[\xi]_i$ and $[\eta]_k$ be paths descending from a vertex a to vertices b and c respectively. Let π_1 be the path $-\xi]_i + [\eta]_k$, and let it be assumed that paths $\pi_2, \pi_3, \dots, \pi_r$, each leading from a to b , have been defined. Let X_r be the (finite) indexed subcomplex of Σ formed by all the cells occurring in the paths π_1, \dots, π_r . The *depth* in X_r of a vertex x is defined to be the maximum possible number of cells in a descending path from a to x in X_r , (or 0 if there is no such path). Thus the depth of any vertex in X_{i+1} is not less than its depth in X_i .

If π_r contains no peak, π_{r+1} is not defined. If it contains at least one peak, choose one, say y , of minimum depth in X_r among peaks of π_r . Let the vertices

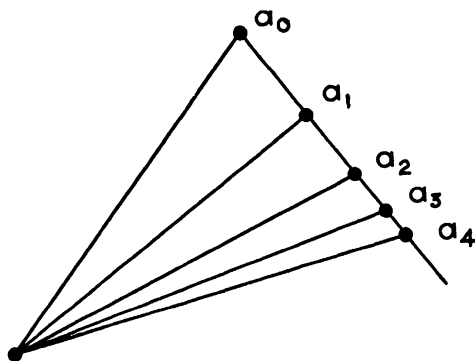


FIG. 1

immediately preceding and succeeding y on π_r be u and v , the (positive) cells yu and yv being ω and ω' respectively. There exist, by (D), paths σ and τ , (either or both of which may be null), descending from u and v to a common vertex w , and π_{r+1} is formed from π_r by substituting $\sigma - \tau$ for $-\omega + \omega'$. The effect is to replace the peak at y by at most two new ones, of depths in X_{r+1} at least 1 greater than that of y in X_r (or zero). By a simple induction it follows that π_r has at most r peaks; and if we make the inductive hypothesis that the peaks of π_{2^n} are of depth at least n in X_{2^n} it follows that, if $r \leq 2^n$, at most $2^n - r$ peaks of π_{2^n+r} are of depth n or less in X_{2^n+r} . Thus the induction is complete and it is proved that if $r \geq 2^n$ all peaks of π_r are of depth at least n .

If $[\zeta]_n$ is a descending path of maximum length in X_r from a to a given vertex z , where $r \geq 2^m$, ζ_i belongs to X_{2^i} for $i = 1, 2, \dots, m$. Suppose that, for a certain i , X_j is the first of the X 's to contain ζ_i , where $j > 2^i$. Then $[\zeta]_i$ is a descending path in X_j of maximum length to its final vertex, z_i , since any longer one could

be used as part of a longer descending path to z in X_r . Thus the depth of z_i in X_j is i . Since z_i belongs to X_j but to no earlier X , it is a cell of one of the descending paths that eliminate a peak, y , in the formation of π_j from π_{j-1} . By the result of the preceding paragraph, if the depth of y in X_{j-1} is p , $j-1 < 2^{p+1}$. But the depth of z_i in X_j exceeds that of y in X_{j-1} by at least 2, $i \geq p+2$. Therefore $j-1 < 2^{i-1}$, $j < 1 + 2^{i-1} \leq 2^i$, contrary to the hypothesis.

The series of paths π_1, π_2, \dots , terminates. If not choose, for each n , a maximal descending path, σ_n , in X_{2^n} from a to a peak of π_{2^n} . Since the first cell of each of these paths is in the finite complex X_2 there is at least one cell, ω_1 , which is the first cell of σ_n for an infinity of n . Since the second cell of each of this infinite subsequence is in X_4 there is at least one cell, ω_2 , such that $\omega_1 + \omega_2$ is the beginning of an infinity of the σ_n . Continuing in this way we obtain an infinite descending path $\omega_1 + \omega_2 + \dots$ in Σ ,—contrary to its given property.

Thus the series of paths π_r from b to c terminates in a path π_q , which, since it has no peak, must descend or ascend directly from b to c , or else descend from b to a vertex w and then rise to c .

The finiteness condition imposed on descending paths in Theorem 3 cannot be replaced by the corresponding "completeness" condition, that every descending chain of vertices has a lower bound in Σ . This is shown by the complex in Fig. 3, in which the vertices c and d are lower bounds of all sets of vertices not containing either of them; but c and d have themselves no lower bound.

6. Topology of Σ

The complex Σ can be made into a 2-complex, Σ^2 , by adding a 2-cell bounded by each of the 1-cycles $\omega + \sigma - \tau - \omega'$ occurring in the proof of theorem 3 (one for each π_r). *Every component of Σ^2 is simply connected.* Any two paths, π and π' , connecting vertices a_1 and b_1 are deformable, by the method of Theorem 3, into paths $\sigma_1 - \tau_1$ and $\sigma'_1 - \tau'_1$ respectively, where σ_1 and τ_1 descend to a vertex a_2 , σ'_1 and τ'_1 to b_2 ; and if \sim stands for "is deformable into," $-\tau_1 + \tau'_1 \sim \sigma_2 - \tau_2$, and $-\sigma_1 + \sigma'_1 \sim \sigma'_2 - \tau'_2$, where $\sigma_2, \tau_2, \sigma'_2, \tau'_2$ are descending paths, the first two to a_3 , the second two to b_3 . In this way paths σ_n and τ_n descending to a_{n+1} , and σ'_n and τ'_n to b_{n+1} , are defined for every n . If an infinity of different paths descending from a_1 could be made from the $\sigma_i, \tau_i, \sigma'_i$ and τ'_i , an infinity of them would necessarily contain one or other of σ_1, σ'_1 ,—say σ_1 ; and of these an infinity would contain one or other of σ_2, σ'_2 ,—say σ'_2 ; and so on. The descending path $\sigma_1 + \sigma'_2 + \dots$ so constructed would have an infinity of different paths as subsets, and would therefore be infinite, contrary to the postulated property of Σ . The number of different paths must therefore be finite.

It follows that for some m , $\sigma_m = \tau_m = \sigma'_m = \tau'_m = 0$; i.e.

$$\pi - \pi' \sim \sigma_1 - \tau_1 - \tau'_1 - \sigma'_1 \sim \sigma_m - \tau_m + \tau'_m - \sigma'_m = 0.$$

7

To establish our second criterion, we suppose that Σ is the sum of two sub-complexes,⁵ L and R , and shall use the terms " L -cell," " R -path," etc., in an obvious sense. " $x\lambda y$ " and " $x\rho y$ " shall mean that there is a cell xy in L or R respectively, and xLy and xRy that $x > y$ in L or R . (In diagrams the positive L - and R -cells will slope down towards the left and right respectively.) We denote by Q the following property of Σ .

(Q) If $x\lambda y$ and $x\rho z$ there exists a vertex w such that zLw , and either $y = w$ or yRw . (We require zLw , which is not necessarily implied by $z = w$. The possibility that $y = z$ is not excluded.)

THEOREM 4. If, in a complex with the property Q , all L -paths are finite, then

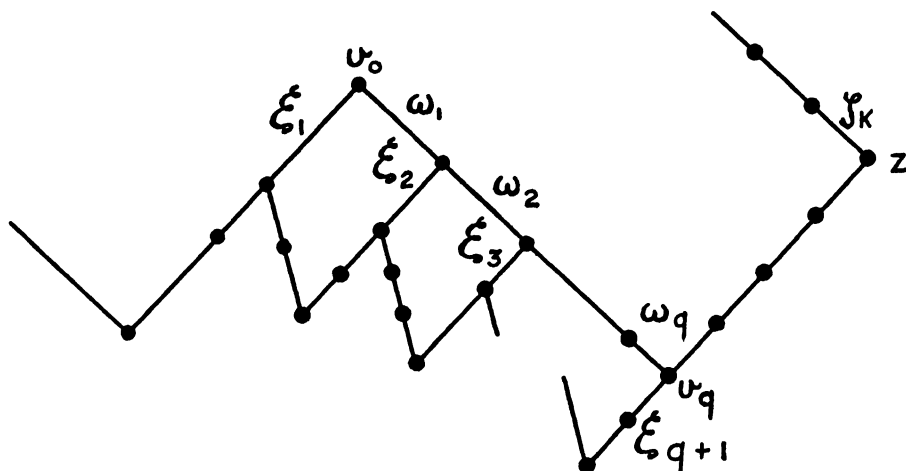


FIG. 2

if xLy and xRz there exists a vertex w such that zLw and either $y = w$ or yRw . If all L -paths are finite, then, in property (Q), $z \neq w$.

It is sufficient to prove the theorem when $x\lambda y$, the general case then following by induction. Let η be an L -cell from x to y , and $[\zeta]_k$ an R -path from x to z . Let π_1 be $-\eta + [\zeta]_k$, and suppose, inductively, that a path π_r from y to z has already been defined.

If π_r has no peak π_{r+1} is not defined; otherwise let v_0 be the last peak on π_r , from y towards z . We assume, inductively, that in proceeding from y towards z the direct ("downward") cells of π_r are in R and the reversed cells in L ,—an assumption evidently satisfied for $r = 1$. The part $v_0 \cdots z$ of π_r is of the form $[\omega]_q - \sigma$ where $[\omega]_q$ is a descending R -path and σ a descending L -path. σ may be null, but $q \neq 0$ since v_0 is a peak. Let ξ_1 be the predecessor of ω_1 in π_r ,

⁵ This always means "indexed subcomplex," the positive direction in each 1-cell agreeing with that in Σ .

(and therefore an L -cell). Assuming inductively that ξ_m is defined, for some $m \leq q$, as an L -cell with the same initial vertex as ω_m , let σ_m and σ'_m be the R - and L -paths which, by (Q), descend from the final vertices of ξ_m and ω_m to a common vertex. Then σ'_m is not null, and we define ξ_{m+1} to be its first cell: say $\sigma'_m = \xi_{m+1} + \tau_m$. The path π_{r+1} is now defined to be the result of substituting $\sigma_1 - \tau_1 + \sigma_2 - \tau_2 + \cdots + \sigma_q - \sigma'_q$ for $-\xi_1 + [\omega]_q$. It evidently has the property that reversed cells are in L and direct cells in R , and the inductive definition of π_r is therefore completed.

If v_q is the final vertex of ω_q , and $-\sigma_r^*$ is the portion $v_q \cdots z$ of π_r , σ_r^* is a descending L -path from z . The corresponding portion of $-\pi_{r+1}$ is $\sigma_r^* + \sigma'_q$, with at least one more cell. Since all L -paths are finite it follows that the process of constructing paths π_r terminates after a certain number, k , of steps, i.e. π_k has no peaks and is therefore a descending (possibly null) R -path from y to a

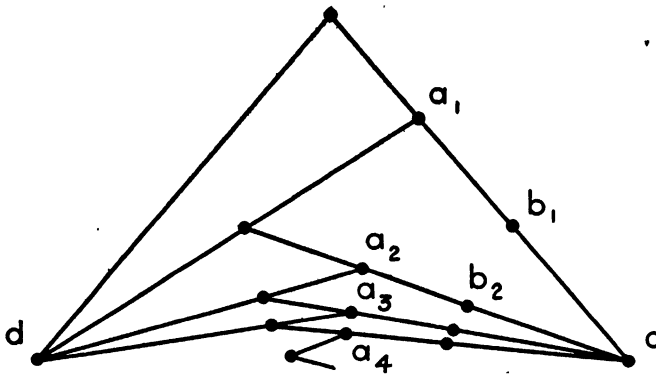


FIG. 3

vertex w , followed by an ascending (non-null) L -path from w to z . Thus yRw or $y = w$, and zLw .

COROLLARY 4.1. *If (Q) is strengthened by excluding the possibility $y = w$, Theorem 4 may be strengthened in the same way. (Obvious from the method of proof.)*

COROLLARY 4.2. *A descending l -path and a descending R -path have at most one common vertex. If the two paths have their initial and final vertices, a and b in common, i.e. if aLb and aRb , there is a vertex c such that bLc and bRc (the alternative $b = c$ being impossible in this case); and a vertex d such that cLd and cRd ; and so on. The path $a \cdots b \cdots c \cdots d \cdots$ is an infinite descending L -path.*

In particular a cell cannot be both an L - and an R -cell. This does not mean that the condition (Q) could be weakened in Theorem 4 by adding "if $y \neq z$ " at the beginning. That this would make the theorem untrue is shown by the example in Fig. 3, where segments sloping down towards the left and right belong to L and R respectively, and the cells marked b,c are in both L and R .

The condition (Q) is satisfied for pairs with $y \neq z$, and no descending L -path has more than two cells; but c and d have no lower bound.

Theorem 4 also fails if the alternative " $z = w$ " is allowed in (Q). This is seen by omitting the vertices b_i in Fig. 3 so that each $a_i c$ becomes a single R -cell (but not now an L -cell).

8

We return to the consideration of indexed 1-complexes in general. A set of (positive) cells, E_x , is at x if x is the initial vertex of every member of E_x . (The null-set is at every vertex.) We suppose that if ξ is a cell xz , each cell η at x has a finite set of cells at z , called $\eta | \xi$, assigned to it as its ξ -derivate.⁶ The ξ -derivate, $E_x | \xi$, of a set E_x is the logical sum of the ξ -derivates of members of E_x . (An appearance of the symbol $E_x | \xi$ implies that ξ is at x .) If π is a descending path from x , $E_x | \pi$, the derivate of E_x by continuation along π , is defined inductively by the equation

$$E_x | (\pi + \xi) = (E_x | \pi) | \xi;$$

and if π is null, $E_x | \pi = E_x$. We usually write $E | (\pi + \xi)$ without brackets: $E | \pi + \xi$. The path $[\xi]_m$ is a development of E_x if, for $1 \leq i \leq m$, $\xi_i \in E_x | [\xi]_{i-1}$. The development is complete if $E_x | [\xi]_m = 0$, partial if not.

The letters CD are used as an abbreviation for "complete development." We postulate the following conditions on the derivates:

- (Δ_1) $\eta | \xi$ is null if, and only if, $\eta = \xi$;
- (Δ_2) if $\eta \neq \zeta$, $(\eta | \xi) \cap (\zeta | \xi) = 0$;
- (Δ_3) if η and ζ are distinct cells at x , there exist developments κ_η and κ_ζ of $\eta | \zeta$ and $\zeta | \eta$ respectively, with a common final vertex w .
- (Δ_4) with the notation of (Δ_3), $\xi | (\eta + \kappa_\zeta) = \xi | (\zeta + \kappa_\eta)$, for any ξ at x .

It follows from (Δ_4), by summation, that the derivates of any set E_x by continuation along $\eta + \kappa_\zeta$ and $\zeta + \kappa_\eta$ are the same. A further consequence is that κ_η and κ_ζ are complete developments of $\eta | \zeta$ and $\zeta | \eta$ respectively. For

$$\begin{aligned} (\eta | \zeta) | \kappa_\eta &= \eta | \zeta + \kappa_\eta \\ &= \eta | \eta + \kappa_\zeta = 0. \end{aligned}$$

From (Δ_2) it follows by induction on the length of π that if $E_x^1 \cap E_x^2$ is null, $(E_x^1 | \pi) \cap (E_x^2 | \pi)$ is also null.

LEMMA 1. If π is a development of E_x^1 , and $E_x^2 | \pi \subseteq E_x^1 | \pi$, then $E_x^2 \subseteq E_x^1$.

Let π be $[\xi]_m$. Let j be the least integer such that $E_x^2 | [\xi]_j \subseteq E_x^1 | [\xi]_j$. If the lemma is false $j \geq 1$, and $E_x^2 | [\xi]_{j-1}$ contains a cell ζ not in $E_x^1 | [\xi]_{j-1}$. Hence $\zeta \neq \xi_j$, and $\zeta | \xi_j$ is a non-null subset of $E_x^2 | [\xi]_j$ not contained in $E_x^1 | [\xi]_j$, contrary to the hypothesis.

In particular if $E_x^2 | \pi = 0$, $E_x^2 \subseteq E_x^1$.

It is assumed, further, that a relation J holds between certain of the pairs of

⁶ For an illustrative example of derivates see §13.

cells at a vertex, and a set E_a is defined to be a J -set if $\xi J \eta$ and $\eta J \xi$ for every distinct pair ξ, η in E_a . (Thus all sets with less than two members are J -sets.)

(J₁) If $\xi J \eta$, $\xi \mid \eta$ has precisely one member.

(J₂) If $\eta_1 \in \xi_1 \mid \zeta$ and $\eta_2 \in \xi_2 \mid \zeta$, and if $\xi_1 J \xi_2$ or $\xi_1 = \xi_2$, then $\eta_1 J \eta_2$ or $\eta_1 = \eta_2$. It follows from J₂ that if E is a J -set, $E \mid \zeta$, and more generally $E \mid \pi$, is a J -set. From J₁ it follows that for no ξ does $\xi J \xi$.

It is now agreed that a set denoted by E, E_a , etc., shall be finite. (A CD of any set is finite by definition.)

LEMMA 2. If the J -set E has k members, all CD's of E have k cells and the same final vertex, and all partial developments are parts of CD's.

If η and ξ are in E , $\eta \mid \xi$ has one member if $\eta \neq \xi$, and none if $\eta = \xi$. Thus $E \mid \xi$ is a J -set with $k - 1$ members, and the development comes to an end after k steps.

Let $\eta + \sigma$ and $\zeta + \tau$ be any two CD's of E , ending at y and z respectively; and let $\eta \neq \xi$. By J₁ and Δ_3 the sets $\eta \mid \zeta$ and $\zeta \mid \eta$ are single cells, η' and ζ' , with a common final vertex, w . By Δ_4 , $E \mid \eta + \zeta' = E \mid \zeta + \eta'$, a set with $k - 2$ members. Let π be a CD of this set, ending at u . By an inductive hypothesis $\zeta' + \pi$ and σ , being CD's of $E \mid \eta$, a J -set with $k - 1$ members, have the same final vertex: $u = y$. Similarly $u = z$, and so $y = z$.

LEMMA 3. If E_x is a J -set, and E_x^1 any set at the same vertex x , all derivatives of E_x^1 by continuation along CD's of E_x are identical.

With the notation of the previous lemma,

$$\begin{aligned} E_x^1 \mid \eta + \sigma &= E_x^1 \mid \eta + \zeta' + \pi, \text{ (inductive hypothesis),} \\ &= E_x^1 \mid \zeta + \eta' + \pi, \text{ (\Delta}_4\text{),} \\ &= E_x^1 \mid \zeta + \tau, \text{ (inductive hypothesis).} \end{aligned}$$

9

If E_x^1 is a J -set, $E_x \mid E_x^1$ denotes the continuation of E_x along a CD of E_x^1 . By Lemma 3 it is independent of the CD chosen. $E_x \mid E_x^1 + E_x^2 + \cdots + E_x^k$ is defined inductively to be $(E_x \mid E_x^1 + \cdots + E_x^{k-1}) \mid (E_x^k \mid E_x^1 + \cdots + E_x^{k-1})$. Thus if $[\xi]_k$ is a CD of E_x^1 , $E_x \mid E_x^1$ and $E_x \mid [\xi]_k$ are alternative notations for the same set.

We now come to the main results of the paper. All the conditions Δ and J are purely local, and involve only a fixed number of given cells.

THEOREM 5. Let π_1 and π_2 be paths in a 1-complex with the properties J₁₋₃ and Δ_1 - Δ_4 descending from a to vertices b and c . Then there exist paths π_3 and π_4 descending from b and c to a common vertex d , such that if E_a is a set at a ,

$$E_a \mid \pi_1 + \pi_3 = E_a \mid \pi_2 + \pi_4.$$

We first prove the following special case.

LEMMA 5.1. If E_a^1 and E_a^2 are J -sets, the CD's of $E_a^1 \mid E_a^2$ and $E_a^2 \mid E_a^1$ have the same final vertex, and if E_a is any set at a , $E_a \mid E_a^1 + E_a^2 = E_a \mid E_a^2 + E_a^1$.

(Since all sets called " E ", " E^1 ", etc., in the following proof are at a , the suffix a will be omitted.)

⁷CASE 1. $E^1 \cap E^2 = 0$. Let $n(E)$ denote the number of elements in E . We proceed by induction on m , $= n(E^1 | E^2) + n(E^2 | E^1)$. Excluding the trivial case where one set E^i is null, the minimum possible value of m is 2. This minimum is only attained if $n(E^1) = n(E^2) = 1$, and Lemma 5.1 then follows from Δ_3 and Δ_4 . We may therefore assume that $m > 2$, and also that $n(E^1) > 1$ or $n(E^2) > 1$, —say $E^1 = E^3 \cup \eta$, where $E^3 \neq 0$.

The proof depends on the fact that if ξ is not in E , $n(E | \xi) \geq n(E)$; and hence if E^p and E^q are J -sets satisfying $E^p \cap E^q = 0$, and $E^p \subseteq E^q$, then $n(E^p | E^q) \leq n(E^p | E^q)$. Thus $n(E^2 | E^3) \leq n(E^2 | E^1)$ and $n(E^3 | E^2) < n(E^1 | E^2)$. Therefore, by the inductive hypothesis, CD's of $E^2 | E^3$ and $E^3 | E^2$ have the same final vertex, z . Since E^1 is a J -set, $\eta | E^3$ is a single cell, ξ , and

$$\begin{aligned} \xi | (E^2 | E^3) &= \eta | E^3 + E^2 = \eta | E^2 + E^3 & (\text{by the inductive hypothesis}) \\ &= E^1 | E^2 + E^3, \end{aligned}$$

since $E^3 | E^2 + E^3 = E^3 | E^3 + E^2 = 0$. Thus there is a CD of $E^1 | E^2$ consisting of a CD of $E^3 | E^2$ followed by a CD of $\xi | (E^2 | E^3)$. Since $E^3 | E^2$ is not null it follows that $n(\xi | (E^2 | E^3)) < n(E^1 | E^2)$, and since also $(E^2 | E^3) | \xi = E^2 | E^1$ the inductive hypothesis may be applied to the sets ξ and $E^2 | E^3$ at z . The final vertices of CD's of $(E^2 | E^3) | \xi$, i.e. $E^2 | E^1$, and of $\xi | (E^2 | E^3)$ are therefore identical, and the latter set has been seen to be the end portion of a CD of $E^2 | E^1$. The first part of the induction is therefore complete. If E is any set at a ,

$$\begin{aligned} E | E^2 + E^1 &= E | E^2 + E^3 + \eta, \\ &= E | E^3 + E^2 + \eta, & (\text{by the inductive hypothesis applied to } E^2 \text{ and } E^3) \\ &= E | E^3 + \eta + E^2, & (\text{by the inductive hypothesis applied to } \xi \text{ and } E^2 | E^3) \\ &= E | E^1 + E^2. \end{aligned}$$

CASE 2. As CASE 1 save that $E^1 \cap E^2 \neq 0$. Let $E^i = E^0 \cup E^{i+2}$, for $i = 1, 2$, where $E^3 \cap E^4 = 0$. By CASE 1, applied to $E^4 | E^0$ and $E^3 | E^0$, $E^4 | E^0 + E^3$ and $E^3 | E^0 + E^4$ have the same final vertex, and since $E^0 | E^0 = 0$ these two sets are $E^2 | E^0 + E^3$ and $E^1 | E^0 + E^4$, i.e. $E^2 | E^1$ and $E^1 | E^2$.

In the *general case*, to which we now turn, the result may be stated more explicitly as follows, taking π_1 and π_2 to be $[\eta]_j$ and $[\xi]_k$.

LEMMA 5.2. *If $[\eta]_j$ and $[\xi]_k$ are any paths descending from a , to b and c there exist paths σ_r and τ_s , (possibly null, $r = 1, \dots, j+1$, $s = 1, \dots, k+1$) such that*

- (1) $\eta_s = \sigma_{1s}$, $\xi_r = \tau_{r1}$,
- (2) $\sigma_{r+1,s}$ and $\tau_{r,s+1}$ have the same final vertex,

⁷ This proof of Case 1 was suggested by Dr. J. H. C. Whitehead, in place of one based on Theorem 4.

(3) σ_{rs} is a CD of E_{rs}^1 , $= \eta_s | \tau_{1s} + \tau_{2s} + \cdots + \tau_{r-1,s}$, and τ_{rs} of E_{rs}^2 , $= \zeta_r | \sigma_{r1} + \cdots + \sigma_{r,s-1}$.

(4) for any E_a , $E_a | \pi_1 + \tau_{1,k+1} + \cdots + \tau_{i,k+1} = E_a | \pi_2 + \sigma_{i+1,1} + \cdots + \sigma_{i+1,k}$.

Starting from the two given paths $[\eta]_i$ and $[\zeta]_k$ we add, one by one, the pairs of paths $\sigma_{r+1,s}$ and $\tau_{r,s+1}$ for the couples (r, s) in the standard "triangular" order $(1, 1), (1, 2), (2, 1), (1, 3), \dots$. When the time comes for $\sigma_{r+1,s}$ and $\tau_{r,s+1}$ to be added, the paths σ_{rs} and τ_{rs} , descending from a vertex x_{rs} , and corresponding to the earlier couples $(r, s-1)$ and $(r-1, s)$, have already been constructed as CD's of E_{rs}^1 and E_{rs}^2 . Hence by the cases of Theorem 5 already settled, CD's $E_{rs}^1 | \tau_{rs}$ and $E_{rs}^2 | \sigma_{rs}$, i.e. of $E_{r,s+1}^1$ and $E_{r+1,s}^2$, meet at a common vertex. These CD's are $\tau_{r,s+1}$ and $\sigma_{r+1,s}$; the induction is complete. (In the limiting cases $r = 1$ and $s = 1$ the single cells η_r and ζ_s play the parts of E_a^1 and E_a^2 in the earlier cases.)

The proof just given provides a method of deforming $\pi_1 + \pi_3$ into $\pi_2 + \pi_4$ by a series of steps in each of which a path $\tau_{rs} + \sigma_{r+1,s}$ is replaced by a path $\sigma_{rs} + \tau_{r,s+1}$. By Lemma 5.1 a set E_x at the common initial vertex of σ_{rs} and τ_{rs} , when continued along either of these paths gives the same result, and therefore the continuation of E_a along the whole path is unaffected by a single step.

THEOREM 6. Any two CD's of a (finite) set E_x have the same final vertex.

If $[\eta]_i$ and $[\zeta]_k$, ending in b and c , are the developments then, with the notations of Lemma 5.2, since $\eta_s \in E_x | [\eta]_{s-1}$,

$$\begin{aligned} E_{rs}^1 &\subseteq E_x | [\eta]_{s-1} + \tau_{1s} + \cdots + \tau_{r-1,s}, \\ &= E_x | [\zeta]_{r-1} + \sigma_{r1} + \cdots + \sigma_{r,s-1}, \end{aligned}$$

and therefore $\sigma_{r1} + \sigma_{r2} + \cdots$ is a development of $E_x | [\zeta]_{r-1}$; and in particular $\sigma_{k+1,1} + \sigma_{k+1,2} + \cdots$ is a development of $E_x | [\zeta]_k$, $= 0$, since $[\zeta]_k$ is a CD. Thus $c = d$, and similarly $b = d$.

COROLLARY 6.1. Continuation of a set E_a along any two CD's of a set E_a^1 gives the same result. This now follows from Theorem 5, π_3 and π_4 being null.

Corollary 6.1 cannot be extended to give the general monodromy property, "continuation of E_a along any two descending paths from a to b gives the same result." Consider the 1-complex in Fig. 5, in which the vertices marked x are identical. Derivates are defined by parallel displacement downward, except that the derivates of xy and xz at z and y are zw and yw respectively. All sets are J -sets. The conditions Δ and J are satisfied in this complex, but continuation of ab to x via b gives the null set, via c the cell xz .

THEOREM 7. In a complex satisfying J and Δ , all developments of a finite set E_x are finite.

Every set is a sum of J -sets, namely its individual members. We proceed by induction on the smallest number, k , of J -sets, E_x^k , whose sum is the given set E_x . (The case $k = 1$ is Lemma 2.)

There is at least one CD of E_x , namely $[\sigma]_k$, where σ_r is a CD of the J -set $E_x^r \mid [\sigma]_{r-1}$. Suppose that $\zeta_1 + \zeta_2 + \dots$ is an infinite development of E_x , and let σ_{rs} , τ_{rs} , E_{rs}^1 and E_{rs}^2 be as in Lemma 5.1, save that σ_r replaces η_r . Then just as in Theorem 6, $\tau_{1s} + \tau_{2s} + \dots$ is a development of $E_x \mid [\sigma]_{s-1}$. Since, for $i < k$, E_x^i is annihilated by continuation along σ_i , $E_x \mid [\sigma]_{k-1} = E_x^k \mid [\sigma]_{k-1}$.

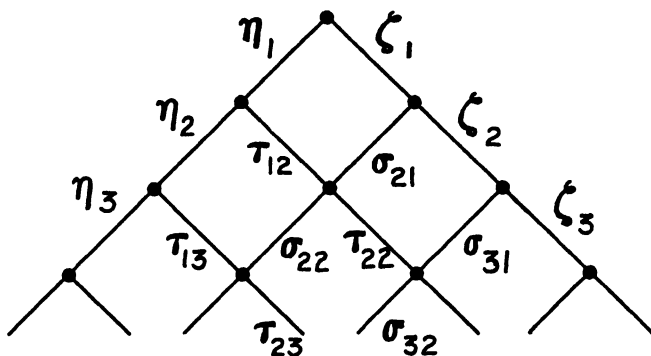


FIG. 4

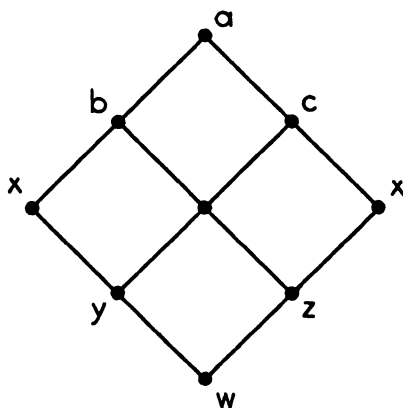


FIG. 5

Thus $\tau_{1k} + \tau_{2k} + \dots$ is a development of a J -set, and so $\tau_{rk} = 0$ if r exceeds a certain q . Therefore if $r > q$,

$$0 = E_{rk}^2 = \zeta_r \mid \sigma_{r1} + \dots + \sigma_{r,k-1}.$$

Now $\sigma_{r1} + \dots + \sigma_{r,k-1}$ is a development of $H_r = (E_x^1 \cup E_x^2 \cup \dots \cup E_x^{k-1}) \mid [\zeta]_{r-1}$, and hence by Lemma 1, $\zeta_r \in H_r$. Therefore the infinite path $\zeta_{q+1} + \zeta_{q+2} + \dots$ is a development of H_{q+1} , a sum of $(k-1)$ J -sets,—contrary to the inductive hypothesis.

COROLLARY 7.1. *There are only a finite number of different developments of E_x .*

If there are an infinity, some cell ξ_1 of the finite set E_x must come first in an

infinity of developments; and some cell ξ_2 of the finite set $E_x \mid \xi_1$ must be second in an infinity of these developments; and so on. The path $\xi_1 + \xi_2 + \dots$ is an infinite development of E_x , contrary to Theorem 7.

10

Theorem 7 is connected with the problem of "random reduction". To a "normal form" or "end-form" in a system with moves there corresponds an end of Σ , and to a normal form of X an end connected by a path to a given vertex x . It follows from Theorem 5 that there is a descending path from x to the end, and that a vertex cannot be connected to two ends, i.e. that an "object" in the corresponding system cannot have two different normal forms. There remains, however, the possibility of an infinite descending path from a vertex which is also connected to an end. It will now be shown that this possibility is not realised in complexes satisfying the conditions Δ and J.

THEOREM 8. *If, in a complex satisfying the conditions Δ and J, there is a path descending from x to an end e of Σ , all descending paths from x are finite, and all maximal paths end at e .*

That all maximal descending paths from x end at e is obvious in view of Theorem 5; only the finiteness remains to be proved.

Let $[\eta]_m$ be a descending path from x to e , and (if possible) $\xi_1 + \xi_2 + \dots$ an infinite descending path from x . Let the paths σ_r and τ_r , and the sets E_r^1 and E_r^2 , be constructed as in Lemma 5.2. Since e is an end all the $\tau_{r,m+1}$ are null. Let j be the largest number such that $\tau_{r,j}$ is non-null for an infinity of values of r , and k a number such that $\tau_{r,j+1} = 0$ if $r \geq k$. Then $E_{r,j+1}^2$, of which $\tau_{r,j+1}$ is a CD, is also null, giving $E_{r,j}^2 \mid \sigma_{r,j} = E_{r,j+1}^2 = 0$. Since $\sigma_{r,j}$ is a CD of $E_{r,j}^1$, it follows (Lemma 1) that, for $r \geq k$,

$$E_{r,j}^2 \subseteq E_{r,j}^1 = E_{k,j}^1 \mid \tau_{k,j} + \dots + \tau_{r-1,j}.$$

Thus $\tau_{k,j} + \tau_{k+1,j} + \dots$ is a development of $E_{k,j}^1$, and by Theorem 7 cannot be infinite,—contrary to the definition of j .

It follows that if a 2-complex Σ^2 is constructed as in §5, all its components containing ends of Σ are simply connected.

11

The theorems that have been proved indicate that complications will arise when the descending paths that join the "y" and "z" of condition (D) to "w" have either more or less than one member each, and that the difficulties are of a different kind in the two cases. In the foregoing group of theorems the second possibility, (corresponding to $\xi \mid \eta = 0$ for $\xi \neq \eta$), was excluded. The following theorem allows this possibility, but is in other ways more special than Theorem 5, and the meaning of the conditions imposed is less obvious. The theorem is used in extending the Church-Rosser Theorem to an enlarged calculus.

We suppose that derivates are defined in Σ , and satisfy Δ_2 - Δ_4 , but that Δ_1 holds only in the weakened form

$$(\Delta_1^*) \quad \xi \mid \xi = 0, \text{ and if } \xi \mid \eta = 0 \text{ then } \eta A \xi,$$

where $\eta A \xi$ stands for "either $\eta = \xi$ or $\eta \mid \xi$ has just one member". The following additional "J-condition" is imposed, \bar{A} denoting "not A ":

(J₃) if $\xi \bar{A} \eta$ and $\xi J \zeta$, then $\eta \bar{J} \zeta$.

(Condition J₃ does not imply the second half of Δ_1^* , since $\xi \bar{J} \xi$.)

Lemmas 2 and 3 remain true under these conditions, and are proved as before. The notation $E_a \mid E_a^1 + \cdots + E_a^k$ may therefore be introduced for J -sets E_a^i .

THEOREM 9. A complex with the properties Δ_1^* , $\Delta_2 - \Delta_4$ and J₁-J₃ has the property (A).

It is sufficient to prove the following special case, since the extension to the general case then proceeds exactly as in Theorem 5.

LEMMA 9.1. If E_a^1 and E_a^2 are J -sets, and E_a is any set at a , the CD's of $E_a^1 \mid E_a^2$ and $E_a^2 \mid E_a^1$ have the same final vertex, and $E_a \mid E_a^1 + E_a^2 = E_a \mid E_a^2 + E_a^1$.

Let $[\eta]_k$ and $[\zeta]_k$ be CD's of E_a^1 and E_a^2 , η_i and ζ_i having final vertices b_i and c_i . " $E_a^1 J E_a^2$ " means " $\eta J \zeta$ if $\eta \in E_a^1$ and $\zeta \in E_a^2$." From J₂ it follows that if $E_a^1 J E_a^2$, $(E_a^1 \mid \xi) J (E_a^2 \mid \xi)$.

CASE 1: $E_a^1 J E_a^2$. We show further that in this case $E_a^1 \mid E_a^2$ has j cells. First let $j = 1$. If also $k = 1$ the result follows immediately from J₁ and Δ_3 . For

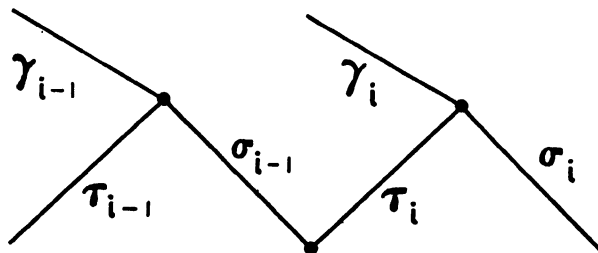


FIG. 6

general k , we have $(\eta_1 \mid \zeta_1) J (E_a^2 \mid \zeta_1)$; and since $\eta_1 \mid \zeta_1$ is a single cell, and $E_a^2 \mid \zeta_1$ has $k - 1$ cells, an inductive hypothesis shows that $\eta_1 \mid E_a^2$ is a single cell and has the same final vertex as a CD, π' , of $(E_a^2 \mid \zeta_1) \mid (\eta_1 \mid \zeta_1)$, which by Δ_4 is $E_a^2 \mid \eta_1 + \zeta_1$. Hence π' is a CD of $E_a^2 \mid \eta_1$. That $E_a \mid E_a^1 + E_a^2 = E_a \mid E_a^2 + E_a^1$ is proved, as in Theorem 5, by repeated applications of Δ_4 . Case 1 for general j is now completed by applying the case $j = 1$ successively to η_r and $E_a^2 \mid [\eta]_{r-1}$, for $r = 1, 2, \dots, j$, and using the last part of the result for $r - 1$.

CASE 2: $E_a^1 \mid E_a^2 = 0$. We show further that, in this case, $E_a^2 \mid E_a^1$ has k members or less. If $j = k = 1$ the result is clear from J and Δ . Suppose that $j = 1, k > 1$. Then $\xi_1 A \eta_1$, for if $\xi_1 \bar{A} \eta_1$, by J₃ and J₂ $(\eta_1 \mid \zeta_1) J (E_a^2 \mid \zeta_1)$, and hence by Case 1 $\eta_1 \mid E_a^2 \neq 0$, contrary to the hypothesis. Thus $\xi_1 \mid \eta_1$ is a single cell. By a k -induction applied to $\eta_1 \mid \zeta_1$ and $E_a^2 \mid \zeta_1$, (in place of E_a^1 and E_a^2), all CD's of $E_a^2 \mid \zeta_1 + \eta_1$ have $k - 1$ cells or less, and end at c_k . Since this set is also $E_a^2 \mid \eta_1 + \zeta_1$, the CD of $E_a^2 \mid \eta_1$ is the cell $\xi_1 \mid \eta_1$, followed by the $k - 1$ cells, (or less), of $E_a^2 \mid \zeta_1 + \eta_1$. The final part follows by repeated applications of Δ_4 .

The extension to general j is as in Case 1.

GENERAL CASE. In view of Lemma 3 it may be assumed that if $E_a^1 | E_a^2 \neq 0$, $[\eta]_i$ is chosen so that $\eta_1 | E_a^2 \neq 0$. A series of "zig-zag" paths, π_1, π_2, \dots , from b_i to c_k , is constructed, π_1 being $-\eta]_i + [\zeta]_k$. Suppose π_r already constructed,

$$\pi_r = \tau_0 - \sigma_1 + \tau_1 - \dots + \tau_{m-1} - \sigma_m,$$

where σ_i and τ_i are CD's of subsets, E_i^1 and E_i^2 , of $E_a^1 | \gamma_i$ and $E_a^2 | \nu_i$ respectively. The γ_i are descending paths from a satisfying

(i) γ_m is $[\zeta]_k$

(ii) for any E_a at a , $E_a | \gamma_i + \sigma_i = E_a | \gamma_{i-1} + \tau_{i-1}$.

This whole inductive hypothesis is satisfied by π_1 if $m = 2$, $\tau_0 = \sigma_2 = \gamma_1 = 0$, $\sigma_1 = \gamma_0 = [\eta]_i$, $\tau_1 = \gamma_2 = [\zeta]_k$.

Let π_r have a peak at the join of σ_{m-1} and τ_{m-1} , and let u be the final vertex of τ_{m-1} . First suppose that $E_{m-1}^1 | \tau_{m-1} = 0$. By Case 2 a CD, θ , of $E_{m-1}^2 | \sigma_{m-1}$ ends at u , and we define π_{r+1} to be $\tau_0 - \sigma_1 + \dots + \tau_{m-2} + \theta - \sigma_m$, the new m' , τ'_{m-2} and σ'_{m-1} being $m-1$, $\tau_{m-2} + \theta$ and σ_m . The γ_i up to γ'_{m-2} are the corresponding γ_i , and $\gamma'_{m-1} = \gamma_m$. Since θ is a CD of

$$\begin{aligned} E_{m-1}^2 | \sigma_{m-1} &\subseteq E_a^2 | \gamma_{m-1} + \sigma_{m-1} \\ &= E_a^2 | \gamma_{m-2} + \tau_{m-2}, \end{aligned}$$

$\tau_{m-2} + \theta$ is a CD of a subset of $E_a^2 | \gamma_{m-2}$. For any E_a at a ,

$$\begin{aligned} E_a | \gamma'_{m-1} + \sigma'_{m-1} &= E_a | \gamma_m + \sigma_m \\ &= E_a | \gamma_{m-1} + \tau_{m-1} \\ &= E_a | \gamma_{m-1} + \sigma_{m-1} + \theta \quad (\text{Case 2}) \\ &= E_a | \gamma_{m-2} + \tau'_{m-2}. \end{aligned}$$

Secondly let $E'_{m-1} | \tau_{m-1} \neq 0$. By Lemmas 2 and 3, if a CD of E'_{m-1} , whose first cell, $\xi^{(1)}$, satisfies $\xi^{(1)} | \tau_{m-1} \neq 0$, is substituted for σ_{m-1} , all the conditions imposed on π_r remain satisfied, and it may therefore be assumed that σ_{m-1} itself is such a CD. Let τ_{m-1} be $[\omega]_p$, ($p \neq 0$ in view of the peak). We construct successively, as in Theorem 5, for $i = 1, 2, \dots, p$, pairs of descending paths τ'_{m-2+i} and $\xi^{(i+1)} + \sigma'_{m-1+i}$, which are CD's of $\omega_i | \xi^{(i)}$ and $\xi^{(i)} | \omega_i$ respectively, and, by Δ_3 , have a common final vertex. The notation " $\xi^{(i+1)} + \sigma'_{m-1+i}$ " implies that $\xi^{(i)} | \omega_i \neq 0$, which is justified by $\xi^{(i-1)} | \omega_{i-1} \neq 0$, derived ultimately from $\xi^{(1)} | \tau_{m-1} \neq 0$. If σ_{m-1} is $\xi^{(1)} + \sigma'_{m-1}$, π_{r+1} is defined to be

$$\tau_0 - \sigma_1 + \dots + \tau_{m-2} - \sigma'_{m-1} + \tau'_{m-1} - \dots + \tau'_{m-2+p} - \sigma'_{m-1+p} - \xi^{(p+1)} - \sigma_m.$$

Its final ascending part has at least one more cell than σ_m . If γ'_h is taken to be γ_h for h up to $m-2$, $\gamma_{m-1} + [\omega]_{h-m+1} + \xi^{(h-m+2)}$ for $h = m-1$ to $m+p-2$, and $\gamma'_{m+p-1} = \gamma_m$, all the conditions are fulfilled, the new " m " being $m+p-1$,

the new " σ_m " $\sigma_m + \xi^{(p+1)} + \sigma'_{m+p-1}$. Condition (ii) follows for the γ'_i immediately from Δ_4 except for $i = m + p - 1$; and

$$\begin{aligned} E_a | \gamma_m + \sigma_m + \xi^{(p+1)} + \sigma'_{m+p-1} &= E_a | \gamma_{m-1} + \tau_{m-1} + \xi^{(p+1)} + \sigma'_{m+p-1} \\ &= E_a | \gamma_{m-1} + [\omega]_{p-1} + \xi^{(p)} + \tau'_{m+p-2} \quad (\Delta_4) \\ &= E_a | \gamma'_{m+p-2} + \tau'_{m+p-2}, \end{aligned}$$

as required.

It has thus been shown how, in all cases where π_r has a peak, a path π_{r+1} is to be constructed having either one less peak or a longer final ascending portion. Since this portion is a development of the finite J -set $E'_a | [\xi]_k$, the second alternative can only occur a finite number of times. Thereafter the number of peaks decreases at each step until a path with no peaks is reached.

The extension to paths which are not CD's of J -sets now follows as in Theorem 5.

12

The theorems that follow Theorem 5, as far as Corollary 6.1, are proved under the new conditions with only minor changes in the argument. Theorem 8 fails to survive, as may be seen by considering Fig. 1: all the conditions Δ_1^* , Δ_2 - Δ_4 , and J_1 - J_3 are satisfied by taking the derivate of a_b at a_{r+1} to be $a_{r+1}b$, and that of $a_r a_{r+1}$ at b to be null; and $a_r b J a_r a_{r+1}$ but not $a_r a_{r+1} J a_r b$.

Theorem 7 still holds but its proof needs some modification.

THEOREM 10. *In a complex satisfying Δ_1^* , Δ_2 - Δ_4 and J_1 - J_3 all developments of finite sets are finite.*

We proceed as in the proof of Theorem 7. As before it follows that $\tau_{1s} + \tau_{2s} + \dots$ is a development of $E_x | [\sigma]_{s-1}$ and that for some p and q the development $\tau_{p+1,q} + \tau_{p+2,q} + \dots$ of $E_x | [\sigma]_{q-1} + \tau_{1q} + \dots + \tau_{pq}$, $= E_x | \theta_{pq}$ say, is infinite, while $\tau_{r,q+1} = 0$ if $r > p$. Since the σ 's and τ 's are CD's of J -sets it follows from Case 2 of Lemma 9.1 that the number of cells in σ_{rq} is (for $r > p$) non-increasing with increasing r . If ζ_r belongs to $E_x^q | [\xi]_{r-1}$, τ_{rq} is contained in $E_x^q | [\xi]_{r-1} + \sigma_{r1} + \dots + \sigma_{rq} = E_x^q | \theta_{r-1,q}$; by the last part of Lemma 9.1 the path σ_{rq} is a CD of this set, and may therefore be chosen in the form $\tau_{rq} + \sigma'_{rq}$, and $\sigma_{r+1,q}$ to be σ'_{rq} . Thus $\sigma_{r+1,q}$ has less cells than σ_{rq} unless $\tau_{rq} = 0$. If, for $r > r_0$, $\tau_{rq} = 0$ whenever $\zeta_r \in E_x^q | [\xi]_{r-1}$, $\tau_{r_0+1,q} + \dots$ is an infinite development of the sum of the $k - 1$ J -sets $E_x^i | \theta_{r_0q}$ ($i \neq q$), contrary to the inductive hypothesis. Hence the number of cells in σ_{rq} eventually diminishes to zero, and from this point on the τ_{rq} coincide with the $\tau_{r,q+1}$ and are null,—contrary to the initial hypothesis.

COROLLARY 10.1. *Under the same conditions, the number of different developments of E_x is finite. (Compare Corollary 7.1).*

13. Application to the conversion calculus

The formalism first considered is that of Theorems 1 and 2 of Church and Rosser [1], but modified in two ways,—first by the adoption of the simpler bracketing of Church [1] secondly by the entire exclusion of “singular” formulae, i.e. those having “accidental” coincidences between the bound variables.⁸ The WFF’s are therefore rows of the symbols λ , variables, and round brackets, built up according to the following rules: (1) x is a WFF, (2) if M is a WFF containing x as a free variable, $(\lambda x M)$ is a WFF, (3) if A and B are WFF’s whose common variables are free in both, (AB) is a WFF. (A variable is *bound* in any row of symbols in which one of its occurrences immediately succeeds a λ , otherwise *free*.)

The allowed transformations that concern us are

I. To replace each specimen of a bound variable x in X by y , a letter not occurring in X .

The result, Y , of any series of applications of I to X will be called an *adjusted copy* of X , and X conv.-I Y .

II. To replace a part $((\lambda x M)N)$ of X by the result of substituting adjusted copies of N for the specimens of x in M , the new bound variables being all different from each other and those of X .

It is agreed that a WFF denoted by one of the letters U, V , is of the form $((\lambda x M)N)$, and we accordingly speak of “the move U on X ,” or “the move (X, U) ,” if U is a part⁹ of X , meaning the application of Rule II in which U is the part operated on. If Y is the WFF that thus replaces X , we write $(X, U) \rightarrow Y$ and X conv.-II Y . If a series of moves I that turns X into Y turns its part U into V , (Y, V) is an adjusted copy of (X, U) .

To define the *residuals* of a part V of X after the move $((\lambda x M)N)$, suppose that each pair of brackets in M is provided with a numerical suffix, which is left unchanged in applying rule II, and that V is enclosed in the pair ${}_1()$. If the move $((\lambda x M)N)$ turns X to Y , then

- (a) if $V = ((\lambda x M)N)$, V has no residual in Y ;
- (b) if V is a part of N its residuals are the corresponding parts of the adjusted copies of N that replace x in M ;
- (c) in all other cases the residual of V is the part ${}_1()$ of Y .

The complex Σ to which our general theorems will be applied has as a typical vertex the class $[X]$ of all adjusted copies of a WFF X . A positive 1-cell is the class $[(X, U)]$, or briefly $[X, U]$, consisting of (X, U) and all its adjusted copies; and its initial and final vertices are $[X]$ and $[Y]$, where $(X, U) \rightarrow Y$. If V is also a part of X , the $[X, U]$ -derivate of $[X, V]$ consists of all the cells $[Y, V_i]$, where the V_i are the residuals of V in Y . Finally “ $[X, U]J[X, V]$ ” means that (i) neither

⁸ Cf. Newman [2] §3. After the general theoretical work the calculus may be extended, for practical convenience, by re-admitting the singular formulae and resuming the original rules I and II; and it can be shewn without difficulty that (1) every singular WFF X conv.-I a non-singular X' , and (2) if X conv.-II Y , X conv.-I X' , Y conv.-I Y' in the extended calculus, X' and Y' being non-singular, then X' conv.-I-II Y' in the restricted calculus.

⁹ Defined as in Church [1].

\mathbf{U} nor \mathbf{V} is a part of the other, and (ii) the free variables of \mathbf{U} and \mathbf{V} are the same. Thus J is a symmetrical relation, and is independent of the WFF chosen to represent $[\mathbf{X}]$.

With these definitions the conditions J_1 , J_2 and Δ_1 - Δ_4 are satisfied. (J_1): no comment is necessary. (J_2): let η_1 and η_2 be determined by the parts \mathbf{V}_1 and \mathbf{V}_2 of \mathbf{X} , and ξ by \mathbf{U} , $= ((\lambda \mathbf{xM})\mathbf{N})$. If $\mathbf{V}_1 = \mathbf{V}_2$, distinct members of $\eta_1 \mid \xi$ are determined by different adjusted copies of \mathbf{V}_1 , and evidently satisfy (i) and (ii). If \mathbf{V}_1 and \mathbf{V}_2 are not identical they are mutually exterior, and a residual of \mathbf{V}_1 could only be a part of a residual of \mathbf{V}_2 if \mathbf{V}_1 were a part of \mathbf{N} , and \mathbf{V}_2 a part of \mathbf{M} containing \mathbf{x} . This contravenes the condition (ii) for \mathbf{V}_1 and \mathbf{V}_2 since \mathbf{x} is free in \mathbf{M} and cannot occur in \mathbf{N} . A part of \mathbf{X} and its residuals in \mathbf{Y} have the same free variables, except that \mathbf{x} is replaced at all occurrences by adjusted copies of \mathbf{N} . Hence if \mathbf{V}_1 and \mathbf{V}_2 have the same free variables their residuals in \mathbf{Y} have also.

In considering the conditions Δ , let ξ , η , ζ be determined by the moves \mathbf{U}_i on \mathbf{X} , ($i = 1, 2, 3$), where $\mathbf{U}_i = \iota((\lambda \mathbf{x}_i \mathbf{M}_i) \mathbf{N}_i)$, and $(\mathbf{X}, \mathbf{U}_i) \rightarrow \mathbf{Y}_i$. Thus \mathbf{U}_i is the part $\iota()$ of \mathbf{X} .

Δ_1 : no comment is necessary. Δ_2 : suppose that $\mathbf{U}_2 \neq \mathbf{U}_3$. The residuals are clearly distinct if they are determined by the original brackets, or one by old and the other by new brackets. The remaining possibility is that a residual of \mathbf{U}_2 is a part, \mathbf{U}'_2 of an adjusted copy of \mathbf{N}_1 , and a residual of \mathbf{U}_3 is either a different part of the same copy, or part of a different copy,—in any case different from \mathbf{U}'_2 . Δ_3 : the condition is obviously satisfied unless one of \mathbf{U}_2 , \mathbf{U}_3 is part of the other,—say \mathbf{U}_2 of \mathbf{U}_3 . If \mathbf{U}_2 is in \mathbf{M}_3 the residual of \mathbf{U}_2 in \mathbf{Y}_3 , and of \mathbf{U}_3 in \mathbf{Y}_2 , are determined by their original brackets, and since \mathbf{N}_3 contains no copy of \mathbf{x}_2 (a bound variable of \mathbf{M}_3), the order of performance of $\iota()$ and $\iota_3()$ is indifferent. If \mathbf{U}_2 is in \mathbf{N}_3 the final effect is the same whether $\iota_2()$ is performed first on \mathbf{N}_3 , followed by $\iota_3()$, or the residuals of $\iota_2()$ on the adjusted copies of \mathbf{N}_3 in \mathbf{Y}_3 . Δ_4 : Let \mathbf{W} be the final result of either series of moves on \mathbf{x} : it has been shown to be unique to within I-adjustment, and therefore determines a unique vertex, w , of Σ . If \mathbf{U}_1 (or \mathbf{U}_2) is not part of either of the other \mathbf{U}_i 's, its performance, before or after \mathbf{U}_2 (or \mathbf{U}_1), does not affect the residual of \mathbf{U}_3 . We may therefore suppose that one of the \mathbf{U}_i 's contains the other two. If \mathbf{U}_3 is not part of either \mathbf{N}_1 or \mathbf{N}_2 the residual of \mathbf{U}_3 in \mathbf{W} by either route is $\iota_3()$. We therefore assume that \mathbf{U}_1 contains both \mathbf{U}_2 and \mathbf{U}_3 , and that \mathbf{U}_3 is part of either \mathbf{N}_1 or \mathbf{N}_2 . Finally, if both \mathbf{U}_2 and \mathbf{U}_3 are in \mathbf{N}_1 , the same residuals of \mathbf{U}_3 are evidently obtained whether \mathbf{U}_2 is performed on \mathbf{N}_1 before \mathbf{U}_1 , or the corresponding moves on the copies of \mathbf{N}_1 after \mathbf{U}_1 . There remains only the case where \mathbf{U}_2 is part of \mathbf{M}_1 , and \mathbf{U}_3 of either, (α) , \mathbf{N}_1 , or, (β) , \mathbf{N}_2 . (α) : the residuals of \mathbf{U}_3 by either route are the corresponding parts of the adjusted copies of \mathbf{N}_1 that replace \mathbf{x}_1 in the move $\iota_1()$ on \mathbf{Y}_2 . (β) : if the residuals of \mathbf{U}_3 in \mathbf{Y}_2 are the parts enclosed in the brackets $\iota_{11}()$, $\iota_{12}()$, \dots , the residuals in \mathbf{W} by either route are the parts enclosed in the same brackets.

The conditions for all the Theorems 5 to 8 are therefore satisfied, and we obtain the following results.

COROLLARY 11.1. *If \mathbf{X} conv.-I-II \mathbf{Y} and \mathbf{X} conv. I-II \mathbf{Z} , there is a WFF \mathbf{W} such that \mathbf{Y} conv. I-II \mathbf{W} and \mathbf{Z} conv. I-II \mathbf{W} .*

COROLLARY 11.2. *There are only a finite number of different developments of a given set of moves II on a WFF \mathbf{X} . All of them are finite, and all end in adjusted copies of the same WFF.*

COROLLARY 11.3. *A WFF has (apart from I-adjustments) at most one normal form, and if one exists all series of moves II terminate in this normal form or an adjusted copy.*

14

Two generalizations of these theorems were given by Church and Rosser in their paper. The first, to the formalism extended so as to include the δ -symbol, is of no interest in the present connection: it is easily shown that the original conditions Δ and J are still satisfied, and hence that the Corollaries 11 hold. The second generalization (of which the proof was not given by Church and Rosser) is to the formalism in which $(\lambda \mathbf{xM})$ is counted a WFF even if \mathbf{x} does not occur in \mathbf{M} . The rules of procedure, and the definitions of derivates need no modification, and the conditions Δ_2 - Δ_4 are proved to hold, just as before. The second part ("only if") of condition Δ_1 now fails, but the condition Δ_1^* is satisfied. The conditions J_1 - J_3 are also satisfied if a different, more complicated, interpretation is given to J .

Let " $\mathbf{U} S \mathbf{V}$ " stand for "a free variable of \mathbf{U} is bound in \mathbf{V} ." It implies that \mathbf{U} is a proper part of \mathbf{V} , and if \mathbf{U}' and \mathbf{V}' are, for any \mathbf{W} , \mathbf{W} -residuals of \mathbf{U} and \mathbf{V} , $\mathbf{U}' S \mathbf{V}'$ implies $\mathbf{U} S \mathbf{V}$. Let " $\mathbf{U} Ex \mathbf{V}$ " stand for "neither \mathbf{U} nor \mathbf{V} is a part of the other." Then, with the same notation, $\mathbf{U}' Ex \mathbf{V}'$ implies $\mathbf{U} Ex \mathbf{V}$. We now take $[\mathbf{X}, \mathbf{U}] J [\mathbf{X}, \mathbf{V}]$, for any significant \mathbf{U} and \mathbf{V} , to mean

"(i) \mathbf{U} is not a part of \mathbf{V} , and (ii) there is no part \mathbf{W} of \mathbf{X} such that $\mathbf{V} S \mathbf{W}$ and $\mathbf{U} Ex \mathbf{W}$."

J_1 is clearly satisfied.

J_2 . Let the notations be those of the previous discussion of J_2 , and let \mathbf{V}'_1 and \mathbf{V}'_2 be distinct residuals of \mathbf{V}_1 and \mathbf{V}_2 . If $\mathbf{V}'_1 S \mathbf{V}'_2$ and $\mathbf{V}'_2 Ex \mathbf{W}'$, then $\mathbf{V}_1 S \mathbf{W}$ and $\mathbf{V}_2 Ex \mathbf{W}$, which is incompatible with $\eta_1 = \eta_2$ or $\eta_1 J \eta_2$. If $\mathbf{V}_1 = \mathbf{V}_2$, \mathbf{V}'_1 cannot be part of \mathbf{V}'_2 . If $\mathbf{V}_1 \neq \mathbf{V}_2$, the only possibility that \mathbf{V}'_1 be part of \mathbf{V}'_2 is that \mathbf{V}_2 be part of \mathbf{M} , with \mathbf{x} as a free variable, and \mathbf{V}_1 be in \mathbf{N} ,—which in view of $\eta_1 J \eta_2$ contravenes (ii).

J_3 . Let η , ξ and ζ be determined by \mathbf{U} , \mathbf{V}_1 , and \mathbf{V}_2 , where \mathbf{U} is $((\lambda \mathbf{xM})\mathbf{N})$. Suppose that $\xi \bar{A} \eta$ and $\eta J \zeta$. Then \mathbf{V}_1 is part of \mathbf{N} and either \mathbf{U} is part of \mathbf{V}_2 or, for some \mathbf{W} , $\mathbf{V}_2 S \mathbf{W}$ and $\mathbf{U} Ex \mathbf{W}$. The first alternative gives \mathbf{V}_1 part of \mathbf{V}_2 ; the second $\mathbf{V}_2 S \mathbf{W}$ and $\mathbf{V}_1 Ex \mathbf{W}$; and both contradict $\xi J \zeta$. Hence

THEOREM 12. *Corollaries 11.1, 11.2 and the first part of Corollary 11.3 hold in the extended calculus.*

It is easily seen that the second part of Corollary 11.3 fails to survive.

REFERENCES

- J. W. ALEXANDER, [1] *Combinatorial theory of complexes*, Annals of Math., 31 (1930), pp. 292-320.
- G. BIRKHOFF, [1] *Lattice theory*, New York, 1940.
- A. CHURCH, [1] *A formulation of the simple theory of types*, J. of Symbolic Logic, 5 (1940), pp. 56-68.
- A. CHURCH AND B. ROSSER, [1] *Some properties of conversion*, Trans. Amer. Math. Soc., 39 (1936), pp. 472-482.
- M. H. A. NEWMAN, [1] *A theorem in combinatory topology*, J. London Math. Soc., 6 (1931), pp. 186-192.
- [2] *Stratified systems of logic*. (Forthcoming.)

ISOMORPHISMS OF NORMED LINEAR SPACES¹

BY GEORGE W. MACKEY

(Received August 28, 1941)

Introduction

Following Banach [1, p. 180]² we say that two normed linear spaces X_1 and X_2 are isomorphic if there exists a one-to-one correspondence between their elements which is both a homeomorphism and an algebraic isomorphism; that is, if the spaces are abstractly identical as topological linear spaces. Let R_i ($i = 1, 2$) be the ring of all continuous linear³ transformations of X_i into itself. Eidelheit [2] has shown that if X_1 and X_2 are complete, then X_1 and X_2 are isomorphic if and only if R_1 and R_2 are isomorphic as rings. In this paper we prove two analogous theorems; one involving the lattices of closed linear subspaces of X_1 and X_2 and the other their groups of automorphisms (self isomorphisms). The latter theorem differs a little from the others in that from the isomorphism of the groups of automorphisms it is not concluded that X_1 and X_2 are isomorphic but only that either this is the case or X_1 and X_2 are what we shall call pseudo-reflexive and mutually pseudo-conjugate. In neither theorem do we need to assume anything about completeness, and we use our methods to prove Eidelheit's theorem without this restriction.

In all three theorems the proof of the necessity is trivial and that of the sufficiency involves three main steps. First we use the given isomorphism between the associated algebraic systems to set up a one-to-one linear independence preserving correspondence between the one dimensional subspaces of X_1 and X_2 . Next we show that this correspondence may be defined by a one-to-one linear transformation of all of X_1 into all of X_2 . Finally we prove that the transformation is a homeomorphism. The second and third steps are accomplished in the same way in all three cases. We devote the first section to proving the two fundamental lemmas involved. The first step is accomplished by associating one dimensional subspaces with elements of the algebraic systems in a natural way and showing that the correspondence between one dimensional subspaces set up via the given isomorphism between the algebraic systems has the properties desired. This is carried out by giving algebraic characterizations of certain kinds of elements and sets of elements in the algebraic systems. The difficulty of doing this increases rapidly as we pass from lattices through rings to groups. Accordingly we prove the three theorems in that order in sections II, III, and IV.

¹ Presented to the American Mathematical Society under another title, November 22, 1941.

² The numbers in brackets refer to the bibliography.

³ In this paper linear means additive and homogeneous.

I. Two Fundamental Lemmas

LEMMA A. *If X_1 and X_2 are linear spaces having dimension greater than two and if $A \rightleftharpoons A'$ where $A \subset X_1$ and $A' \subset X_2$ represents a one-to-one correspondence between the one dimensional linear subspaces of X_1 and X_2 respectively which preserves linear independence, then there exists a one-to-one linear transformation T from all of X_1 into all of X_2 such that if A_x is the linear subspace of scalar multiples of x , then $A_{T(x)} = A'_x$ for all x in X_1 .*

Let \bar{x} be any non-zero element in X and let \bar{y} be any non-zero element in $A'_{\bar{x}}$. It is clear that if T exists then $T(\bar{x}) = \lambda\bar{y}$ and may be chosen so that $\lambda = 1$. With this choice of $T(\bar{x})$, T is uniquely determined for all x in X_1 . In fact if $x = \lambda\bar{x}$ we must have $T(x) = \lambda\bar{y}$ and if x and \bar{x} are linearly independent then A'_x and $A'_{\bar{x}-x}$ will be linearly independent whereas A'_x , $A'_{\bar{x}}$, and $A'_{\bar{x}-x}$ will be linearly dependent. Hence any element in $A'_{\bar{x}}$, in particular \bar{y} , will be a unique sum of elements x_1 and y_1 from A'_x and $A'_{\bar{x}-x}$ respectively. Since we require that $T(\bar{x}) = T(x) + T(\bar{x} - x)$, we must have $T(x) = x_1$.

It remains to show that the T so defined is linear. It will obviously then have the other required properties. Let x and y be arbitrary elements of X_1 . Let M be a three dimensional subspace of X_1 containing x , y , and \bar{x} . Let M' be the three dimensional subspace of X_2 spanned by the one dimensional subspaces of the form A' where A is in M . Lemma A for three dimensional spaces is well known. It is simply the theorem of projective geometry⁴ to the effect that a collineation between two real projective planes can be represented analytically by a linear transformation [3, vol. I, p. 190 and vol. II, p. 252]. Thus there is a linear transformation of the desired sort taking M into M' . By the argument of the first paragraph, T as on M must be a constant multiple of this transformation. Therefore if λ and μ are arbitrary scalars, $T(\lambda x + \mu y) = \lambda T(x) + \mu T(y)$ and, since x and y were arbitrary, T is linear.

Before stating and proving Lemma B we make a few preliminary remarks concerning the relation between the "maximal" subspaces of a linear space and the linear functionals defined on the space. We define a maximal subspace as a proper subspace contained in no other proper subspace. It is clear that if z is any element in the complement of a maximal subspace M of a linear space X , then any element in X has a unique representation in the form $x = m + \lambda z$ where m is in M and λ is a scalar. If we make the definition $f(m + \lambda z) = \lambda$, it is easily seen that $f(x)$ is a linear functional which vanishes on M and only on M . Conversely, if $f(x)$ is any non trivial linear functional defined on X , it is easily verified that the set of elements x of X such that $f(x) = 0$ is a maximal subspace of X . We call it the null-space of $f(x)$. Finally, suppose that $f_1(x)$ and $f_2(x)$ are non trivial linear functionals having the same null-space. Let z be in the complement of this subspace. Then $f_2(z)f_1(x) - f_1(z)f_2(x)$ is zero for all x in X . Therefore $f_2(x) = kf_1(x)$ where k is a constant. Thus there is a

⁴ Thanks are due the referee for suggesting the use of this theorem to eliminate a large part of the author's original proof.

natural one-to-one correspondence between the maximal subspaces of X and the one dimensional subspaces of the space of all linear functionals on X . If $f(x)$ is continuous as well as linear then its null space is obviously closed. (We suppose now that X is normed.) Conversely, given any closed maximal subspace M of X , it follows from the lemma on page 57 of [1] that there exists a non-trivial continuous linear functional vanishing on M and hence having M for its null-space. Thus every linear functional having M as its null-space is continuous. In other words our one-to-one correspondence associates closed subspaces with continuous functionals and vice-versa.

LEMMA B. *If X_1 and X_2 are normed linear spaces and T is a one-to-one linear transformation of all of X_1 into all of X_2 such that T and T^{-1} carry maximal closed subspaces into maximal closed subspaces, then T is a homeomorphism and hence X_1 and X_2 are isomorphic.*

Let $f_2(x)$ be an arbitrary non-trivial continuous linear functional defined on X_2 . Let M_2 be the null-space of f_2 and let $M_1 = T^{-1}(M_2)$. M_1 is a closed maximal subspace of X_1 . Therefore there exists a continuous linear functional $f_1(x)$ defined on X_1 , having M_1 for its null-space. Let z be a member of the complement of M_1 . By adjusting the arbitrary scalar multiplier, f_1 may be chosen so that $f_1(z) = 1$. If x is an arbitrary member of X_1 , we may write $x = m + f_1(x)z$ where m is in M_1 . Therefore $f_2(T(x)) = f_2(T(m)) + f_2(f_1(x)T(z)) = f_2(T(z))f_1(x)$ since $T(m)$ is in M_2 . Now let $\{x_n\}$ be any bounded sequence of elements of X_1 . Then $\{f_1(x_n)\}$ is a bounded sequence of real numbers. Hence since $f_2(T'(x_n)) = f_2(T(z))f_1(x_n)$, $\{f_2(T(x_n))\}$ is a bounded sequence of real numbers. But f_2 may be any continuous linear functional on X_2 . Therefore, by [1, p. 80 Théorème 6], $\{T(x_n)\}$ is a bounded sequence of elements of X_2 . Thus T takes bounded sets into bounded sets. Hence using the theorem on page 54 of [1] we conclude that T is continuous. By an exactly analogous argument, T^{-1} is continuous. This completes the proof.

Lemma B essentially says that the topology of a normed linear space is determined as soon as it is given which maximal linear subspaces are closed. This is closely related to a theorem of Fichtenholz [4] to the effect that the topology is determined by the set of continuous linear functionals.

II. The Lattice Theorem

THEOREM. *Let X_1 and X_2 be normed linear spaces. Let L_1 be the lattice of closed linear subspaces of X_1 and L_2 that of X_2 . Then X_1 and X_2 are isomorphic as normed linear spaces if and only if L_1 and L_2 are isomorphic as lattices.*

We begin by proving some lemmas.

DEFINITION. *If S_1 and S_2 are subsets of a linear space we denote by $S_1 \dot{+} S_2$ the smallest linear subspace containing both S_1 and S_2 , and by $S_1 \dot{\vdash}$ the smallest linear subspace containing S_1 .*

LEMMA 2.1. *If M is a closed linear subspace of a normed linear space X and \bar{x} is any element of X , then $M \dot{+} \bar{x}$ is closed.*

Neglecting trivial cases we suppose that \bar{x} is not in M and that $M \dot{+} \bar{x} \neq X$. By the lemma on page 57 of [1], there exists a continuous linear functional f_1 on X which vanishes on all members of M and is such that $f_1(\bar{x}) = 1$. Let y be any element of X such that any continuous linear functional which vanishes on all members of $M \dot{+} \bar{x}$ also vanishes on y . If $y - f_1(y)\bar{x}$ is not in M , then, again by the lemma on page 57 of [1], there exists a continuous linear functional f_2 , vanishing on all members of M and such that $f_2(y - f_1(y)\bar{x}) = 1$. But $f_2 - f_2(\bar{x})f_1$ vanishes on all members of $M \dot{+} \bar{x}$ and hence on y . Therefore $f_2(y) - f_2(\bar{x})f_1(y) = f_2(y - f_1(y)\bar{x}) = 0$. Therefore $y - f_1(y)\bar{x}$ is in M and y is in $M \dot{+} \bar{x}$. Hence if y is any element not in $M \dot{+} \bar{x}$, there exists a continuous linear functional vanishing on all members of $M \dot{+} \bar{x}$ and not vanishing on y . In other words $M \dot{+} \bar{x}$ is an intersection of null-spaces of continuous linear functionals and hence is closed.

As a corollary we have

LEMMA 2.2. *Any finite dimensional subspace of a normed linear space is closed.*

LEMMA 2.3. *If L is the lattice of closed linear subspaces of a normed linear space and n is a positive integer, then a member M_n of L has dimension n if and only if there exist members of L , M_1, M_2, \dots, M_{n-1} , such that M_1 covers the zero of L and M_{i+1} covers M_i for $i = 1, 2, \dots, n-1$.*

If M has dimension n , let x_1, x_2, \dots, x_n be a basis for M . By Lemma 2.2, $x_1 \dot{+}, x_1 \dot{+} x_2, \dots, x_1 \dot{+} x_2 \dot{+} \dots \dot{+} x_n$, are all members of L , and obviously each covers its predecessor; except of course $x_1 \dot{+}$ which covers 0. Conversely, since if N is finite dimensional and N' covers N , it is obvious that the dimension of N' is one greater than that of N , we conclude from the existence of such a chain that M has dimension n .

We turn now to the proof of the theorem. If X_1 and X_2 are isomorphic it is obvious that L_1 and L_2 are isomorphic. Suppose, conversely, that L_1 and L_2 are isomorphic. If A is any one dimensional subspace of X_1 , it follows from Lemma 2.2 that A is in L_1 . Let A' be the correspondent of A in L under the lattice isomorphism. Since A covers 0, A' covers 0 and hence is one dimensional. In this way we set up a one-to-one correspondence between the one dimensional subspaces of X_1 and X_2 respectively. Let A_1, A_2, \dots, A_r be a linearly independent set of one dimensional subspaces of X_1 , and let A'_1, A'_2, \dots, A'_r be their respective correspondents in X_2 . If A'_1, A'_2, \dots, A'_r are linearly dependent, then $A'_1 \dot{+} A'_2 \dot{+} \dots \dot{+} A'_r = M'$ has dimension $k < r$. By Lemma 2.2, M' is in L_2 . Let M be the correspondent of M' in L . It is an easy consequence of Lemma 2.3 that M has the same dimension as M' . But since $A'_i \subset M'$, $A_i \subset M$ ($i = 1, 2, \dots, r$). Therefore $A_1 \dot{+} A_2 \dot{+} \dots \dot{+} A_r$ has dimension less than r . This contradicts our hypothesis that A_1, A_2, \dots, A_r are linearly independent. Since the same argument applies to linearly independent sets of subspaces of X_2 , we see that our one-to-one correspondence preserves linear independence. Hence if neither X_1 nor X_2 has dimension less than three we may apply Lemma A and conclude the existence of a one-to-one linear transformation T of all of X into all of X such that $A_{T(x)} = A'_x$ for all x in X_1 . Let

M be any closed maximal subspace of X_1 . M is covered by \hat{X}_1 . Hence its correspondent in L_2 under the lattice isomorphism, M' , is covered by X_2 and consequently is maximal. Now x is in M if and only if $A_x \subset M$, which, by virtue of the lattice isomorphism, is the case if and only if $A'_x \subset M'$. But $A'_x = A_{T(x)}$. Therefore x is in M if and only if $T(x)$ is in M' . It follows that $T(M) = M'$ and hence that T takes closed maximal subspaces into closed maximal subspaces. Similarly, T^{-1} does likewise and applying Lemma B, we conclude that X_1 and X_2 are isomorphic. Suppose finally that either X_1 or X_2 has dimension less than three. Then since X_1 and X_2 correspond under the lattice isomorphism, it follows from Lemma 2.3 that the other has this same dimension. But any two finite dimensional normed linear spaces having the same dimension are isomorphic by Lemma B, since, by Lemma 2.2, all of the maximal subspaces of both are closed. This completes the proof.

III. The Ring Theorem

DEFINITION. A continuous linear transformation E of a normed linear space X into itself such that $E^2 = E$ will be called a projection. The set of elements of X of the form $E(x)$, where x is in X , will be called the range of the projection E . The dimension of the range will be called the dimension of E . If E_1 and E_2 are projections such that the range of E_1 is contained in the range of E_2 we shall say that E_1 is contained in E_2 , and if furthermore E_2 is not contained in E_1 , we shall say that E_1 is properly contained in E_2 . Following the terminology used in lattice theory, we shall say that a projection E_2 covers a projection E_1 if E_1 is contained properly in E_2 and there exists no projection E_3 contained properly in E_2 and properly containing E_1 .

LEMMA 3.1. Given any finite dimensional subspace M of a normed linear space X , there exists a projection E whose range is M .

Let x_1, x_2, \dots, x_n be a basis for M . For each $i = 1, 2, \dots, n$ set $f_i(c_1x_1 + \dots + c_nx_n) = c_i$. Then $f_i(x)$ is a linear functional defined on M , and since its null-space is finite dimensional and hence closed, it is continuous. By the Hahn-Banach extension theorem, [1, p. 55, Théorème 2], f_i can be extended so as to be a continuous linear functional F_i defined throughout X . Let $E_i(x) = F_i(x)x_i$. Since F_i is continuous and linear, E_i is a continuous linear transformation of X into itself; hence so is $E = E_1 + E_2 + \dots + E_n$. Finally, $E_i(E_j(x)) = F_i(x_j)F_j(x)x_i$. This is identically zero for $i \neq j$ and identically $E_i(x)$ for $i = j$. Therefore $E^2 = (E_1 + E_2 + \dots + E_n)^2 = E_1 + E_2 + \dots + E_n = E$. Therefore E is a projection whose range is obviously M .

LEMMA 3.2. Given any closed maximal subspace M of a normed linear space X , there exists a projection E whose range is M .

Let z be a member of the complement of M . Let f be a linear functional whose null-space is M and choose the arbitrary scalar multiple so that $f(z) = 1$. Set $E(x) = x - f(x)z$. Since M is closed, f is continuous. Therefore $E(x)$ is a continuous linear transformation of X into itself. $f(E(x)) = f(x) - f(x) = 0$. Therefore $E^2(x) = E(E(x)) = E(x) - f(E(x))z = E(x)$. Thus $E(x)$ is a projec-

tion. Finally, since $f(E(x)) \equiv 0$, the range of $E(x)$ is contained in M , and furthermore, since if x is in M , then $f(x) = 0$ so that $E(x) = x$, we see that the range of E is M .

LEMMA 3.3. *If E_1 and E_2 are projections defined on a normed linear space, then E_1 is contained in E_2 if and only if $E_2E_1 = E_1$.*

If $E_2E_1 = E_1$ and x is in the range of E_1 , then $x = E_1y$. But $E_2E_1(y) = E_1(y)$; that is $x = E_2(x)$. Therefore x is in the range of E_2 . Conversely, if E_1 is contained in E_2 , then for each x in X , $E_1(x) = E_2(y)$. Therefore $E_2(E_1(x)) = E_2^2(y) = E_2(y) = E_1(x)$. That is, $E_2E_1 = E_1$.

LEMMA 3.4. *If n is a positive integer and E_n is a projection defined on a normed linear space, then E_n has dimension n if and only if there exist projections E_1, E_2, \dots, E_{n-1} such that E_{i+1} covers E_i for $i = 1, 2, \dots, n-1$, and E_1 covers 0.*

The truth of this lemma follows easily from Lemma 3.1 using an argument similar to that used in proving Lemma 2.3.

LEMMA 3.5. *If E is a projection on a normed linear space X , then the range of E is a closed maximal subspace of X if and only if the projection 1 covers E .*

If the range of E is maximal it is obvious that 1 covers E . Conversely, suppose that 1 covers E . If $x = E(y)$, then $E(x) = E(E(y)) = E(y) = x$. Therefore $x - E(x) = 0$. If $x - E(x) = 0$, then $x = E(x)$. In other words, the range of E is the null-space of the continuous transformation $1 - E$ and hence is closed. It follows from the lemma on page 57 of [1] that the range of E is contained in a closed maximal subspace M of X and hence by Lemma 3.2, there exists a projection E' whose range is M and which contains E . Since 1 covers E , M must also be the range of E .

THEOREM. *Let X_1 and X_2 be normed linear spaces. Let R_1 be the ring of all continuous linear transformations of X_1 into itself and R_2 that of X_2 . Then X_1 and X_2 are isomorphic as normed linear spaces if and only if R_1 and R_2 are isomorphic as rings.*

The necessity is obvious. The proof of the sufficiency is so much like that of the lattice theorem that we shall not give it in detail. Obviously, the ring isomorphism takes projections into projections. From Lemma 3.3 it follows that inclusion of projections is preserved. Combining Lemmas 3.3 and 3.4 we conclude that finite dimensional projections correspond to projections of the same dimension, and using Lemma 3.5 instead of Lemma 3.4, that projections with closed maximal ranges go into projections with closed maximal ranges. We set up a one-to-one correspondence between the one dimensional subspaces of X_1 and X_2 by passing from such a subspace A in X , through a projection E having A for its range, to the range A' of its correspondent E' in R . That A' is uniquely determined by A is an easy consequence of Lemma 3.3. We establish the preservation of linear independence much as we did in the lattice theorem; using Lemma 3.1 to give us a k dimensional projection containing r one dimensional projections whose ranges are linearly dependent. We show that the T we get by using Lemma A is a homeomorphism using Lemma B and Lemma 3.5. Finally, if X_1 or X_2 has dimension less than three, we show that

X_1 and X_2 have the same finite dimension and hence are isomorphic by observing that the units of R_1 and R_2 must correspond under the ring isomorphism and hence being finite dimensional projections must have the same dimension.

Eidelheit's proof of this theorem is considerably shorter than ours. This is principally due to the fact that by using a device apparently only applicable in the ring situation he is able to avoid having to prove Lemma A. We give the longer proof here in order to emphasize the close relationship existing between this theorem and the other two.

IV. The Group Theorem

We begin by discussing the notion of pseudo-reflexivity. Let X be a normed linear space and let \bar{X} be its conjugate space. For each x in X , as is well known, if we define $F_x(f) = f(x)$ for all f in \bar{X} , F_x is a member of \bar{X} and $\|F_x\| = \|x\|$. In general, there will be members of \bar{X} which have no such representation. In the contrary case \bar{X} is said to be reflexive. Even if X is not reflexive it may be such that a new norm may be introduced into \bar{X} under which a linear functional is continuous if and only if it is an F_x . We shall call such a space pseudo-reflexive. The new norm in \bar{X} is not uniquely determined but by virtue of the theorem of Fichtenholz referred to at the end of Lemma B this is the case for the corresponding norm topology. The topological linear space which \bar{X} becomes under this topology we call the pseudo-conjugate of X . Obviously, the pseudo-reflexivity and the pseudo-conjugate of X depend only upon the norm topology in X and not upon the particular norm. Therefore we may speak of the pseudo-conjugate of the pseudo-conjugate of a pseudo-reflexive space X , and it is obvious that it always exists and is isomorphic to X . If X_1 is pseudo-reflexive and X_2 is isomorphic to the pseudo-conjugate of X_1 so that X_2 is also pseudo-reflexive and X_1 is isomorphic to the pseudo-conjugate of X_2 , we say briefly that X_1 and X_2 are pseudo-reflexive and mutually pseudo-conjugate.

A few words on the relationship between reflexivity and pseudo-reflexivity. Clearly reflexive spaces are pseudo-reflexive. On the other hand, we can show without difficulty that if X is complete and pseudo-reflexive, then X is reflexive. In fact if X is pseudo-reflexive and complete, let F be a member of the second conjugate of X . Let $\{f_n\}$ be a sequence of members of \bar{X} , bounded as a sequence of members of the pseudo-conjugate of X . Then $\{f_n(x)\}$ is a bounded sequence of real numbers for each x in X . Therefore, by [1, p. 80, Théorème 5], $\{f_n\}$ is bounded as a sequence of members of the ordinary conjugate of X and hence $\{F(f_n)\}$ is a bounded sequence of real numbers. Therefore F is a continuous linear functional on the pseudo-conjugate of X [1, p. 54, Théorème 1] and hence is an F_x . Thus X is reflexive. Finally, since, as is shown in [5], the conjugate of any normed linear space is complete, no incomplete space is ever reflexive. In other words, a pseudo-reflexive space is reflexive if and only if it is complete. We have examples which we expect to publish later of non-complete pseudo-reflexive normed linear spaces.

THEOREM. *Let X_1 and X_2 be normed linear spaces. Let G_1 be the group of all linear transformations of X_1 into all of itself which are continuous and have continuous inverses and let G_2 be that of X_2 . Then G_1 and G_2 are isomorphic as groups if and only if either (a) X_1 and X_2 are isomorphic as normed linear spaces or (b) X_1 and X_2 are pseudo reflexive and mutually pseudo-conjugate.*

We preface the proof of the theorem proper with a series of lemmas concerning what we shall call involutions.⁵ Given a normed linear space X , a continuous linear transformation of X into itself such that $T^2 = 1$ will be called an involution. Clearly any involution on X is a member of the group G for X . If T is an involution on X , let M_+ be the subspace of X containing all elements in X such that $T(x) = x$, and let M_- be the subspace containing all those such that $T(x) = -x$. Let x be any element in X . We may write $x = \frac{1}{2}(x + T(x)) + \frac{1}{2}(x - T(x))$. But $T(\frac{1}{2}(x + T(x))) = \frac{1}{2}(T(x) + x) = \frac{1}{2}(x + T(x))$ and $T(\frac{1}{2}(x - T(x))) = \frac{1}{2}(T(x) - x) = -\frac{1}{2}(x - T(x))$. Hence x can be represented as the sum of an element in M_+ and an element in M_- . Since M_+ and M_- have nothing in common but 0, this representation is unique. Thus each involution T "decomposes" X into two disjoint closed subspaces in one of which T is 1 and in the other of which T is -1 . These subspaces will be called the subspaces of T . If at least one of them is finite dimensional, the dimension of the one of smaller dimension will be called the dimension of T . If neither is finite dimensional T will be said to be infinity dimensional.

LEMMA 4.1. *If X is a normed linear space, M is a finite dimensional subspace and M' is any closed subspace of X such that $M \dot{+} M' = X$ and $M \cap M' = 0$, then there exists an involution T having M and M' for its subspaces.*

Let x_1, x_2, \dots, x_n be a basis for M . For each $i = 1, 2, \dots, n$, consider $M_i = M' \dot{+} x_1 \dot{+} x_2 \dot{+} \dots \dot{+} x_{i-1} \dot{+} x_{i+1} \dot{+} \dots \dot{+} x_n$. By Lemma 2.1, M is closed and since $M \cap M' = 0$ and the x_i are linearly independent, x_i is in the complement of M_i . Hence there exists a continuous linear functional f_i which has M_i for its null-space and is such that $f_i(x_i) = 1$. Let $T(x) = 2f_1(x)x_1 + 2f_2(x)x_2 + \dots + 2f_n(x)x_n - x$. Then $T(x)$ is a continuous linear transformation of X into itself such that $T(x_i) = 2x_i - x_i = x_i$ ($i = 1, 2, \dots, n$) and $T(x) = -x$ for all x in M' . Therefore $T^2(x) = x$ for all x in X and T is an involution. Since T is 1 in M and -1 in M' it follows readily that M and M' are the subspaces of T .

LEMMA 4.2. *If M is a finite dimensional subspace of a normed linear space X , then there exists an involution T having M as one of its subspaces.*

By Lemma 3.1, there exists a projection E whose range is M . Let M' be the null space of E . Then as is readily verified M and M' satisfy the hypotheses of Lemma 4.1.

LEMMA 4.3. *Let M_+ and M_- be the subspaces of an involution T on a normed linear space X . Let U_+ and U_- respectively be arbitrary continuous linear transformations of M_+ and M_- into themselves. Let U be the unique linear transformation coinciding on M_+ and M_- with U_+ and U_- . Then U is continuous.*

⁵ Cf. Sobczyk [6] page 80.

Let $\{x_n\}$ be a bounded sequence of elements of X . Then $\{x_n + T(x_n)\}$ is a bounded sequence of elements of M_+ and $\{x_n - T(x_n)\}$ is a bounded sequence of elements of M_- . Accordingly, $\{U_+(x_n + T(x_n))\}$ is a bounded sequence of elements of M_+ , and $\{U_-(x_n - T(x_n))\}$ is a bounded sequence of elements of M_- . But $U_+(x_n + T(x_n)) + U_-(x_n - T(x_n)) = U(x_n + T(x_n) + x_n - T(x_n)) = 2U(x_n)$ ($n = 1, 2, \dots$). Therefore $\{U(x_n)\}$ is a bounded sequence of elements of X . Hence by [1, p. 54, Théorème 1], U is continuous.

LEMMA 4.4. *If T is an involution on a normed linear space X and U is an arbitrary linear transformation of X into itself then $UT = TU$ if and only if $U(M_+) \subset M_+$ and $U(M_-) \subset M_-$ where M_+ and M_- are the subspaces of T .*

If $UT = TU$ and x is in M_- then $T(U(x)) = U(T(x)) = U(-x) = -U(x)$. Hence $U(x)$ is in M_- . Similarly, if x is in M_+ then $T(U(x)) = U(T(x)) = U(x)$ and $U(x)$ is in M_+ . Conversely, suppose $U(M_+) \subset M_+$ and $U(M_-) \subset M_-$. If x is in X , then $x = x_+ + x_-$ where x_+ is in M_+ and x_- is in M_- . $UT(x) = U(x_+ - x_-) = U(x_+) - U(x_-)$. $TU(x) = T(U(x_+) + U(x_-)) = U(x_+) - U(x_-)$. Therefore $UT(x) = TU(x)$ for all x in X .

LEMMA 4.5. *If U is a linear transformation of a normed linear space X into itself and $UT = TU$ for every involution T on X , then U is a constant; that is, there exists a scalar λ such that $U(x) = \lambda x$ for all x in X .*

Given any x in X , by Lemma 4.2, there exists an involution T one of whose subspaces is $x \perp$. Since $UT = TU$, it follows from Lemma 4.4 that $U(x) = \lambda_x x$. Let x and y be any two elements of X . $U(x + y) = \lambda_{x+y}(x + y) = \lambda_x x + \lambda_y y$. Hence if x and y are linearly independent, $\lambda_x = \lambda_{x+y} = \lambda_y$. If $y = \mu x$, then $U(y) = \mu U(x) = \mu \lambda_x x = \lambda_x y$. In any case $\lambda_x = \lambda_y$. Therefore there exists λ , independent of x , such that $U(x) = \lambda x$ for all x in X .

DEFINITION. *If A is an arbitrary set of continuous linear transformations of a normed linear space into itself, we shall let A^* denote the set of all involutions T such that $UT = TU$ for every U in A .*

DEFINITION. *Let T_0 be an involution. For each $n = 1, 2, \dots$ choose involutions T_1, T_2, \dots, T_n not necessarily distinct, such that $T_i T_j = T_j T_i$ ($i, j = 0, 1, \dots, n$). For each choice of T_1, T_2, \dots, T_n there will be a certain number of elements in $(T_0, T_1, \dots, T_n)^*$. As we shall see, these numbers are finite and for fixed T_0 and n form a bounded set. We denote by $f(T_0, n)$ the largest number in the set for each T_0 and n .*

LEMMA 4.6. *If T_0 is an involution on a normed linear space X and X is not finite dimensional, then T_0 is finite dimensional if and only if $\sup_n (f(T_0, n)/2^{(2^n)})$ is finite and if T_0 is finite dimensional, its dimension is equal to $\log_2(\sup_n (f(T_0, n)/2^{(2^n)}))$.*

Let T_1, T_2, \dots, T_n be such that $T_i T_j = T_j T_i$ ($i, j = 0, 1, \dots, n$). Denote the subspace on which T_j is 1 by M_j^1 and that on which it is -1 by M_j^{-1} ($j = 0, 1, \dots, n$). Since $T_0 T_1 = T_1 T_0$ it follows from Lemma 4.4 that $T_1(M_0^1) \subset M_0^1$ and $T_1(M_0^{-1}) \subset M_0^{-1}$. Hence $M_0^1 = (M_0^1 \cap M_1^1) \perp (M_0^1 \cap M_1^{-1})$ and similarly for M_0^{-1} . Let $M_0^{(-1)^{i_0}} \cap M_1^{(-1)^{i_1}} = X_{i_0 i_1}$ ($i_j = 0, 1; j = 0, 1$). Then $X = X_{00} \perp X_{01} \perp X_{10} \perp X_{11}$ where each X_{ij} has nothing in common with the \perp union of

all the rest except 0 and we see that T_0 and T_1 are constant in each $X_{i_0 i_1}$ and by applying Lemma 4.3 twice that any linear transformation of X into itself which takes each $X_{i_0 i_1}$ into itself (that is, $T(X_{i_0 i_1}) \subset X_{i_0 i_1}$) and is continuous on each $X_{i_0 i_1}$ is continuous. Since $T_2 T_0 = T_0 T_2$ and $T_2 T_1 = T_1 T_2$, T_2 takes each $X_{i_0 i_1}$ into itself and accordingly decomposes each $X_{i_0 i_1}$ into $X_{i_0 i_1 1}$ and $X_{i_0 i_1 0}$. Continuing this process we finally obtain $X = X_{11 \dots 1} \dot{+} X_{11 \dots 0} \dot{+}$ the summation extending over X 's having as subscripts all $(n+1)$ -uples of 0's and 1's, where $X_{i_0 i_1 \dots i_n} = M_0^{(-1)^{i_0}} \cap M_1^{(-1)^{i_1}} \dots \subset M_n^{(-1)^{i_n}}$. Furthermore, as above, we conclude that each T_j ($j = 0, 1, \dots, n$) is constant in each $X_{i_0 i_1 \dots i_n}$ and that any linear transformation of X into itself which takes each $X_{i_0 i_1 \dots i_n}$ into itself and is continuous on each $X_{i_0 i_1 \dots i_n}$ is continuous. Let U be any member of $(T_0, T_1, \dots, T_n)^*$. Using Lemma 4.4, it is easy to see that U takes each $X_{i_0 i_1 \dots i_n}$ into itself. Conversely, since each T_j is constant in each $X_{i_0 i_1 \dots i_n}$, any such U is in $(T_0, T_1, \dots, T_n)^*$ provided that it is an involution. In other words we get the general member of $(T_0, T_1, \dots, T_n)^*$ by considering an arbitrary involution on each of the subspaces $X_{i_0 i_1 \dots i_n}$ of X and taking the unique linear transformation coinciding with these where they are defined. Since, in particular, we may select the involution 1 in all but one of the $X_{i_0 i_1 \dots i_n}$ and -1 in the one remaining, we see by Lemma 4.4 that any member of $(T_0, T_1, \dots, T_n)^{**}$ must take each $X_{i_0 i_1 \dots i_n}$ into itself. Finally, since given any involution defined in an $X_{i_0 i_1 \dots i_n}$, there exists an involution in $(T_0, T_1, \dots, T_n)^*$ coinciding with the given one where it is defined, it follows from Lemma 4.5 that any member of $(T_0, T_1, \dots, T_n)^{**}$ must be constant on each $X_{i_0 i_1 \dots i_n}$. Conversely, if we assign 1's and -1 's in an arbitrary manner to the $X_{i_0 i_1 \dots i_n}$, it is clear that there exists a unique linear transformation of X into itself which in each $X_{i_0 i_1 \dots i_n}$ is 1 or -1 according to the above assignment and that this transformation is a member of $(T_0, T_1, \dots, T_n)^{**}$. Thus the number of members of $(T_0, T_1, \dots, T_n)^{**}$ is 2^k where k is the number of non-zero $X_{i_0 i_1 \dots i_n}$. This justifies the statement made in the definition of $f(T, n)$ and tells us that $f(T_0, n) \leq 2^{(2^n)} \cdot 2^{(2^n)}$. On the other hand if T is m dimensional where m is a positive integer, then since half of the subspaces $X_{i_0 i_1 \dots i_n}$ are subspaces of an m dimensional space we have $f(T_0, n) \leq 2^m 2^{(2^n)}$. In other words if we make the convention that $2^m = \aleph_0$ whenever T is infinity dimensional then we have in any case $f(T_0, n) \leq 2^{(2^n)} \min(2^m, 2^{(2^n)})$. We shall now show that the equality holds. Suppose first that T_0 is infinity dimensional or that it is m dimensional where m is an integer and $m \geq 2^n$. Then in M_0^1 we may select 2^n linearly independent elements x_1, x_2, \dots, x_{2^n} . By Lemma 4.2, there exists an involution defined on M_0^1 having $x_2 \dot{+} x_3 \dot{+} \dots \dot{+} x_{2^n}$ for one of its subspaces. Let M_1 be the other subspace of this involution. Similarly define $N_1, y_2, y_3, \dots, y_{2^n}$ in M_0^{-1} . Let $M_i = x_i \dot{+}$ and let $N_i = y_i \dot{+}$ ($i = 2, 3, \dots, 2^n$). Then $X = M_1 \dot{+} M_2 \dot{+} \dots \dot{+} M_{2^n} \dot{+} N_1 \dot{+} N_2 \dot{+} \dots \dot{+} N_{2^n}$ where all of the subspaces concerned are closed and at least one dimensional and each has nothing but 0 in common with the $\dot{+}$ union of the rest. Put the subspaces M_1, M_2, \dots, M_{2^n} into one-to-one correspondence with the 2^n $(n+1)$ -uples of 0's and 1's $(0, i_1, i_2, \dots, i_n)$ and the

subspaces N_1, N_2, \dots, N_{2^n} with those of the form $(1, i_1, i_2, \dots, i_n)$. Now given any $j = 1, 2, \dots, n$, consider the unique linear transformation T_j which is 1 in each subspace associated with an (i_0, i_1, \dots, i_n) for which $i_j = 0$ and is -1 on each subspace for which $i_j = 1$. Since $M_1 \perp M_2 \perp \dots \perp M_{2^n} = M_0^1$ and $N_1 \perp N_2 \perp \dots \perp N_{2^n} = M_0^{-1}$ and all of these subspaces except possibly M_1 and N_1 are one dimensional, it follows from Lemmas 2.1, 4.1, and 4.3 that T_j is continuous. Furthermore it is clear that $T_j^2 = 1$ and that $T_i T_j = T_j T_i$ ($i, j = 0, 1, 2, \dots, n$). Finally we see that each $X_{i_0 i_1 \dots i_n}$ for this choice of T_1, T_2, \dots, T_n is the M_i or N_i associated with (i_0, i_1, \dots, i_n) and hence is at least one dimensional. In other words $(T_0, T_1, \dots, T_n)^{**}$ contains $2^{(2^n+1)} = 2^{(2^n)} \cdot 2^{(2^n)} = 2^{(2^n)} \cdot \min(2^{(2^n)}, 2^m)$ members. Suppose now that T_0 is m dimensional where m is an integer less than 2^n . We may suppose without loss of generality that the m dimensional subspace of T_0 is the one on which T_0 is 1. Let x_1, x_2, \dots, x_m be the elements of a basis for M_0^1 . Let $M_i = x_i \perp$ ($i = 1, 2, \dots, m$). Let M_i ($i = m+1, m+2, \dots, 2^n$) denote the 0 dimensional subspace. Now we proceed exactly as before and define involutions T_1, T_2, \dots, T_n such that $T_i T_j = T_j T_i$ ($i, j = 1, 2, \dots, n$) but such that exactly $2^n + m$ of the $X_{i_0 i_1 \dots i_n}$ are at least one dimensional. We are assured of being able to get our full quota of non-trivial N 's by our hypothesis that X is not finite dimensional. Thus in this case also the T_j may be chosen so that $(T_0, T_1, \dots, T_n)^{**}$ contains $2^{(2^n)} \cdot \min(2^{(2^n)}, 2^m)$ members. Hence $f(T_0, n) = 2^{(2^n)} \cdot \min(2^{(2^n)}, 2^m)$. Now consider the behavior of $f(T_0, n)/2^{(2^n)}$ as n increases. If T_0 is infinity dimensional, then $f(T_0, n)/2^{(2^n)} = \min(2^{(2^n)}, 2^m) = 2^{(2^n)}$ which increases without limit. If T_0 is m dimensional where m is an integer, then $f(T_0, n)/2^{(2^n)}$ increases with n until $n \geq \log_2 m$ and thereafter we have $f(T_0, n)/2^{(2^n)} = 2^m$. Therefore $f(T_0, n)/2^{(2^n)}$ is bounded and has 2^m for its greatest value. Thus $m = \log_2 (\sup_n (f(T_0, n)/2^{(2^n)}))$ and the lemma is proved.

LEMMA 4.7. *If X is a normed linear space, X fails to be finite dimensional if and only if for each positive integer n there exist n distinct involutions T_1, T_2, \dots, T_n such that $T_i T_j = T_j T_i$ ($i, j = 1, 2, \dots, n$). If X is finite dimensional and k is the largest positive integer for which there exist k distinct mutually permutable involutions, then $m = \log_2 k$.*

The proof of this lemma is so like certain parts of the proof of Lemma 4.6 that we omit it.

LEMMA 4.8. *Let X be a non finite dimensional normed linear space. Let T_1 and T_2 be one dimensional involutions which do not commute ($T_1 T_2 \neq T_2 T_1$). Let M_1 and M_2 be their respective infinite dimensional subspaces and let ϕ_1 and ϕ_2 be basis elements for their one dimensional subspaces. Then if T is an involution on X , T is contained in $(T_1, T_2)^{**}$ if and only if one subspace of T contains $M = M_1 \cap M_2$ and the other is contained in $N = \phi_1 \perp \phi_2$.*

We begin with a proof of the sufficiency of the condition. Let U be an arbitrary member of $(T_1, T_2)^*$. Let M_+ and M_- be the subspaces of U . By Lemma 4.4, T_1 takes M_+ into M_+ and M_- into M_- and so does T_2 . Hence each of these is constant in one of M_+ and M_- and is one dimensional in the other.

Suppose that T_1 is constant in one of M_+ and M_- and that T_2 is constant in the other. Then $T_1T_2 = T_2T_1$ in each of M_+ and M_- and hence $T_1T_2 = T_2T_1$ contrary to hypothesis. Hence ϕ_1 and ϕ_2 are both contained in the same subspace which we may suppose to be M_+ . We may suppose without loss of generality that $\phi_1 \perp$ and $\phi_2 \perp$ are the subspaces of T_1 and T_2 respectively in which these transformations are 1. Hence $N \subset M_+$ and $M_- \subset M$. Let T be an arbitrary involution whose 1 space is in N and whose -1 space contains M . Then since $T(x) + x$ is always in the 1 space of T and N is in M , we have $U(T(x) + x) = T(x) + x$ for all x in X and this may be written in the form $U(T(x)) = T(x) - U(x) + x$. On the other hand, since $U(x) - x$ is always in M_- and M_- is in M , we have $T(U(x) - x) = x - U(x)$ for all x in X and this takes the form $T(U(x)) = T(x) - U(x) + x$. Comparing these expressions we conclude that $UT = TU$ and hence that T is a member of $(T_1, T_2)^{**}$. To prove the necessity, let T be an arbitrary member of $(T_1, T_2)^{**}$. Given any member y of the complement of N , let $N' = N \perp y$. Since X is infinity dimensional, so is M and hence there exists \bar{m} in M and not in N' . Let f be the unique linear functional defined on $N' \perp \bar{m}$ such that $f(\phi_1) = f(\phi_2) = 0$ and $f(y) = f(\bar{m}) = 1$. Since any maximal subspace of a finite dimensional normed linear space is closed, any linear functional defined on one is continuous. Hence by the Hahn-Banach extension theorem [1, p. 55, Théorème 2], there exists a continuous linear functional F defined on X such that $F(\phi_1) = F(\phi_2) = 0$ and $F(y) = F(\bar{m}) = 1$. Consider the continuous linear transformation $U(x) = x - 2F(x)\bar{m}$. $U(U(x)) = x - 2F(x)\bar{m} - 2F(x)(\bar{m} - 2\bar{m}) = x$. Therefore U is an involution. Furthermore for $i = 1, 2$, $U(T_i(x)) = T_i(x) - 2F(T_i(x))\bar{m}$ and $T_i(U(x)) = T_i(x) - 2F(x)T_i(\bar{m}) = T_i(x) + 2F(x)\bar{m}$. But since $T_i(x) + x$ is in the 1 space of T_i and so in N , $F(T_i(x) + x) = 0$ and $-2F(T_i(x)) = 2F(x)$ for all x in X . It follows that $UT_i = T_iU$ so that U is in $(T_1, T_2)^*$. Hence $UT = TU$. In other words, $T(x) - 2F(T(x))\bar{m} = T(x) - 2F(x)T(\bar{m})$ or $F(T(x))\bar{m} = F(x)T(\bar{m})$ for all x in X . Thus $T(\bar{m}) = \lambda\bar{m}$. Since \bar{m} may be any element of M not in the (at most one dimensional) intersection of N and M , it follows by an argument similar to that used in Lemma 4.5 that T is constant in M , and hence that M is contained in one of the subspaces of T . There is no loss in generality in supposing that it is the -1 space. Thus $\lambda = -1$ and our next to the last equation becomes $F(T(x) + x) = 0$ for all x in X . In other words, the 1 space of T is contained in a subspace containing N and not y . But y was an arbitrary element of $X - N$. Therefore the 1 space of T is contained in N and the lemma is proved.

DEFINITION. If T_1 and T_2 are one dimensional involutions, we say that (T_1, T_2) is a minimal pair if it is impossible to select distinct one dimensional involutions T_3 and T_4 in $(T_1, T_2)^{**}$ so that $(T_3, T_4)^{**}$ is a proper subset of $(T_1, T_2)^{**}$ and $(T_3, T_4)^{**}$ contains an infinite number of members if $(T_1, T_2)^{**}$ does.

LEMMA 4.9. If T_1 and T_2 are one dimensional involutions defined on a non finite dimensional normed linear space X , then T_1 and T_2 have a common subspace if and only if (T_1, T_2) is a minimal pair.

Let $M_1, M_2, M, \phi_1, \phi_2$, and N be defined as in Lemma 4.8. If $M_1 \neq M_2$ and $\phi_1 \dagger \neq \phi_2 \dagger$, let N_3 and N_4 be two distinct one dimensional subspaces of the two dimensional subspace N neither of which is $M_1 \cap N$. Then $M_1 \dagger N_3 = M_1 \dagger N_4 = X$. By Lemma 4.1, there exist involutions T_3 and T_4 such that M_1 is a subspace of both N_3 and N_4 and N_3 and N_4 are respectively their other subspaces. Since $N_3 \neq N_4$ and $M_1 = M_1$, T_3 and T_4 do not commute. Therefore it follows from Lemma 4.8, provided that T_1 and T_2 do not commute, that T_3 and T_4 are in $(T_1, T_2)^{**}$. Again by Lemma 4.8, since M_1 and M_2 are both maximal and $M_1 \neq M_2$, T_2 is not in $(T_3, T_4)^{**}$. Finally since there are an infinite number of ways of choosing a one dimensional subspace of N different from $M \cap N$, $(T_3, T_4)^{**}$ contains an infinite number of members. Therefore if T_1 and T_2 do not commute, (T_1, T_2) is not a minimal pair. If $T_1 T_2 = T_2 T_1$, then it follows from the argument used in Lemma 4.6 that $(T_1, T_2)^{**}$ contains exactly eight members. If we let $T_3 = T_1$ and $T_4 = -T_1$, the same sort of argument tells us that $(T_3, T_4)^{**}$ contains only four members. Therefore in any case (T_1, T_2) is not a minimal pair. Suppose now that $M_1 = M_2 = M$ and $\phi_1 \dagger \neq \phi_2 \dagger$. By Lemma 4.8, if T_3 and T_4 are one dimensional members of $(T_1, T_2)^{**}$ then $M_3 = M_4 = M$ and ϕ_3 and ϕ_4 are in N . Hence if ϕ_3 and ϕ_4 are linearly independent so that $T_1 T_2 \neq T_2 T_1$ and $\phi_3 \dagger \phi_4 = N$, then again by Lemma 4.8, $(T_3, T_4)^{**} = (T_1, T_2)^{**}$. If $\phi_3 \dagger = \phi_4 \dagger$ then $T_4 = -T_3$ and it follows from Lemma 4.6 that $(T_3, T_4)^{**}$ contains only a finite number of members. But by means of the argument used in the first part of the proof it can be shown that $(T_1, T_2)^{**}$ contains an infinity of members. Hence (T_1, T_2) is a minimal pair. If $M_1 \neq M_2$ and $\phi_1 \dagger = \phi_2 \dagger$ the argument is similar. It depends upon the fact that if $M_3 \neq M_4$ and $M_3 \cap M_4 \supset M$ then $M_3 \cap M_4 = M$ and the fact that there are two linearly independent continuous linear functionals vanishing on M . Finally, if $M_1 = M_2$ and $\phi_1 \dagger = \phi_2 \dagger$, then $T_1 = \mp T_2$ and hence $(T_1, T_2)^{**}$ contains only 1, -1 , T_1 , and $-T_1$. Therefore T_1 and $-T_1$ are the only one dimensional involutions present. Thus (T_1, T_2) is a minimal pair and the proof of the lemma is complete.

LEMMA 4.10. *Let X be a non finite dimensional normed linear space. Let m be a positive integer. Then if T is a one dimensional involution on X and T_1 is an m dimensional one, the one dimensional subspace of T is contained in the m (infinity) dimensional subspace of T_1 if and only if there exists an involution T' having the same one dimensional subspace as T and such that $T' T_1 = T_1 T'$ and is an $m - 1(m + 1)$ dimensional involution.*

Let $\bar{M}_m, M_1, \bar{M}_\infty$, and M_∞ denote the subspaces of T and T_1 . We may suppose without loss of generality that \bar{M}_∞ and M_∞ are the -1 subspaces. If $M_1 \subset \bar{M}_m$, let x_1, x_2, \dots, x_m be a basis for \bar{M}_m such that x_1 is in M_1 . By Lemma 4.1, there exists an involution T' whose 1 subspace is M_1 and whose -1 subspace is $\bar{M}_\infty \dagger x_2 \dagger \dots \dagger x_m$. Both T_1 and T' are -1 in \bar{M}_∞ and both are 1 in M_1 . In $x_2 \dagger \dots \dagger x_m$ one is -1 and the other is 1. Thus $T_1 T' = T' T_1$ and is 1 on $\bar{M}_\infty \dagger M_1$ and -1 on $x_2 \dagger \dots \dagger x_m$; that is, is an $m - 1$ dimensional involution. If $M_1 \subset \bar{M}_\infty$ the argument is similar. However,

instead of a basis, we use the existence of an involution on \bar{M}_∞ having M_1 for a subspace. Conversely, suppose that T' has M_1 for its one dimensional subspace and that $T'T_1 = T_1T'$. Then, by Lemma 4.4, T' takes \bar{M}_∞ into \bar{M}_∞ and \bar{M}_m into \bar{M}_m . Hence it is constant on one of \bar{M}_m and \bar{M}_∞ and a one dimensional involution on the other. In other words, M_1 is contained in either \bar{M}_m or \bar{M}_∞ and $T'T_1$ is obviously accordingly either an $m - 1$ or $m + 1$ dimensional involution.

LEMMA 4.11. *Let X be a non finite dimensional normed linear space. Let m be a positive integer. Then if T is a one dimensional involution on X and T_1 is an m dimensional one, the infinity dimensional subspace of T contains the m (infinity) dimensional subspace of T_1 if and only if there exists an involution T' having the same infinity dimensional subspace as T and such that $T'T_1 = T_1T'$ and is an $m + 1(m - 1)$ dimensional involution.*

The proof is analogous to that of Lemma 4.10.

We turn now to the proof of the theorem. If X_1 and X_2 are isomorphic, the isomorphism of G_1 and G_2 is obvious. Suppose that X_1 and X_2 are pseudo-reflexive and mutually pseudo-conjugate. Then as X_2 is isomorphic to \bar{X}_1 under a norm for which the elements of X_1 define the continuous linear functionals on \bar{X}_1 , it will be sufficient to prove that G_1 is isomorphic to G_3 where G_3 is the group of automorphisms of \bar{X}_1 under this norm. In the rest of this discussion wherever the topology of \bar{X}_1 occurs it will be understood to be the one under which it is isomorphic to X_2 . Given any T in G_1 , let $T' = (T^{-1})^*$ where T^* is Banach's conjugate [1, p. 100]. Then since, as is well known, $(T_1T_2)^* = T_2^*T_1^*$ and $(T_1T_2)^{-1} = T_2^{-1}T_1^{-1}$ it follows that $(T_1T_2)' = T_1'T_2'$. If $T' = 1$ for some T , then $f(T^{-1}(x)) = f(x)$ or $f(T^{-1}(x) - x) = 0$ for all f in X and all x in X . Hence [1, p. 55, Théorème 3] $T^{-1}(x) - x = 0$ whence $T = 1$. Next any T' is continuous and hence, since $(T')^{-1} = (T^{-1})'$, has a continuous inverse and is in G_3 . In fact if $\{f_n\}$ is a bounded sequence of elements of \bar{X}_1 , then $\{f_n(T^{-1}(x))\}$ is a bounded sequence of real numbers for all x in X_1 . In other words if $g_n = T'(f_n)$, $n = 1, 2, \dots$ then $\{g_n(x)\}$ is a bounded sequence for all x in X and hence by [1, p. 80, Théorème 6], $\{T'(f_n)\}$ is a bounded sequence of elements of \bar{X}_1 . Hence T' is continuous [1, p. 54, Théorème 1]. Now let U be any member of G_3 . Then if we identify each member of x with the corresponding functional on \bar{X}_1 and repeat the argument we have just given we find that $(U^{-1})^*$ is a member of G_1 and as is easily verified $((U^{-1})^*)^{-1} = U$. Thus the set of elements of the form T' where T is in G_1 is precisely G_3 and $T \rightarrow T'$ is an isomorphism.

Suppose, conversely, that G_1 and G_2 are isomorphic as abstract groups. If either X_1 or X_2 is finite dimensional, the isomorphism of X_1 and X_2 is an easy consequence of Lemma 4.7 and the argument used in the corresponding part of the lattice theorem. We suppose then that neither X_1 nor X_2 is finite dimensional. Let T_1 and T_2 be two one dimensional involutions in G_1 having the same one dimensional subspace N and distinct infinity dimensional ones M_1 and M_2 . Let T'_1 and T'_2 respectively be their correspondents in G_2 . It follows from

Lemma 4.6 that T'_1 and T'_2 are also one dimensional and from Lemma 4.9 that they have a common subspace.

CASE A. T'_1 and T'_2 have the same one dimensional subspace N' .

Since T_1 and T_2 have different infinity dimensional subspaces and hence do not commute, it is clear that T'_1 and T'_2 have different infinity dimensional subspaces. Let T_3 be any other involution having N for one subspace. Since T'_1 and T'_2 have different infinity dimensional subspaces, T'_3 and one of T'_1 and T'_2 have different infinity dimensional subspaces and hence have the same one dimensional subspace. In other words, T'_3 has N' as a subspace. By the same argument if T' is any involution in G_2 having N' as a subspace, then T has N as a subspace. Now let N_1 be any other one dimensional subspace of X . The unique linear functional f defined on $N_1 \perp N$ such that $f(\phi) = f(\phi_1) = 1$ where ϕ_1 and ϕ are non-zero members of N_1 and N respectively is continuous and hence [1, p. 55, Théorème 2] has a continuous linear extension. The null space M_3 of the extension is closed and maximal and contains neither N nor N_1 . By Lemma 4.1, there exists an involution T_3 having M_3 and N for its subspaces and an involution T_4 whose subspaces are M_3 and N_1 . T'_3 has N' for one subspace and has a subspace in common with T'_4 . Since T_4 does not have N for a subspace, T'_4 cannot have N' . Hence T'_3 and T'_4 have the same infinity dimensional subspace. Let T_5 be any involution having N_1 for a subspace. T'_4 and T'_5 have a subspace in common. If they have the same infinity dimensional subspace, then T'_3 and T'_5 and hence T_3 and T_5 have a subspace in common. Since $N_1 \neq N$, it is their infinity dimensional subspace. Hence $T_5 = \pm T_4$. Hence in any case T'_4 and T'_5 have a common one dimensional subspace. In other words it has been shown that if T and U are one dimensional involutions in G_1 , then T and U have the same one dimensional subspace if and only if T' and U' do. From this point on the argument is so much like that used in the ring theorem that we omit it except to say that Lemmas 3.1, 3.2, 3.3, 3.4 and 3.5 are replaced by Lemmas 4.2, 4.1, 4.10, 4.6 and 4.1 respectively and that we conclude that X_1 and X_2 are isomorphic.

CASE B. T'_1 and T'_2 have the same infinity dimensional subspace.

It follows at once from the argument used in Case A that no pair of non-commuting one dimensional involutions in G with a common one dimensional subspace can have correspondents in G with the same one dimensional subspace. Hence whenever T and U have a common one dimensional subspace, T' and U' have a common infinity dimensional subspace, and if we associate with each one dimensional subspace N of X_1 the common infinity dimensional subspace of the correspondents in G_2 of all members of G_1 having N as a subspace, we readily see that we get a one-to-one correspondence between the one dimensional subspaces of X_1 and the closed maximal subspaces of X_2 . Hence if we associate with each closed maximal subspace of X_2 , the one dimensional subspace of continuous linear functionals having it for their null-space we will have a one-to-one correspondence between the one dimensional subspaces of X_1 and \bar{X}_2 respectively. In order to show that this correspondence preserves linear

dependence, we note first that if M is the intersection of a finite number of closed maximal subspaces of a normed linear space X then there exists a finite dimensional subspace \bar{M} such that $M \dot{+} \bar{M} = X$ and $M \cap \bar{M} = 0$, that while \bar{M} is not uniquely determined its dimension is and is equal to the dimension of the space of linear functionals vanishing on M , and that furthermore this dimension does not exceed the number of maximal subspaces involved. It follows at once that the continuous linear functionals f_1, f_2, \dots, f_n are linearly independent if and only if the intersection of their null-spaces contains the infinity dimensional subspace of an $n - 1$ dimensional involution. Let N_1, N_2, \dots, N_n be a set of one dimensional subspaces of X_1 . Let M_1, M_2, \dots, M_n respectively be their corresponding closed maximal subspaces of X_2 . The N_i ($i = 1, 2, \dots, n$) are linearly dependent if and only if there exists an $n - 1$ dimensional involution in G_1 whose finite dimensional subspace contains all of the N_i . Using Lemmas 4.6, 4.10 and 4.11 and involutions having the N_i as subspaces we conclude from this that the N_i are linearly dependent if and only if there exists an $n - 1$ dimensional involution in G_2 whose infinity dimensional subspace is contained in the intersection of the M_i ; that is, if and only if the continuous linear functionals defining the M_i are linearly dependent. Let $V(f)$ be the one-to-one linear transformation from all of \bar{X}_2 into all of X_1 whose existence we can now conclude from Lemma A. Let $\|x\|$ denote the norm of an element x of X_1 and set $\|f\|_1 = \|V(f)\|$ for each element f in \bar{X}_2 . Obviously this defines a norm in \bar{X}_2 under which \bar{X}_2 is isomorphic to X_1 . Let F be a linear functional on \bar{X}_2 which is continuous with respect to the norm $\|\cdot\|_1$. Let L be the null-space of F and let $M = V(L)$. Then M is a closed maximal subspace of X_1 . Let T be an involution in G_1 having M as a subspace. Let N' be the one dimensional subspace of the correspondent of T in G_2 . Using Lemmas 4.6, 4.10 and 4.11 we see that a one dimensional subspace of X_1 is contained in M if and only if the corresponding closed maximal subspace of X_2 contains N' and hence if and only if the corresponding one dimensional subspace of \bar{X}_2 is contained in the common null-space of the elements of N' , regarded as linear functionals on \bar{X}_2 . Thus L is the null-space of an element of N' . It follows that $F(f) = f(x)$ for some x in N' (and hence in X_2) and all f in \bar{X}_2 . Conversely, given any x in X_2 , let T' be a member of G_2 having $x \dot{+}$ as a subspace, let T be the corresponding member of G , and let M be the infinity dimensional subspace of T . Finally let $L = V^{-1}(M)$. Just as before we can show that that L is the null-space of x and hence since L is closed that $F(f) = f(x)$ is continuous. Thus X_2 is pseudo-reflexive and X_1 is isomorphic to its pseudo-conjugate. In other words X_1 and X_2 are pseudo-reflexive and mutually pseudo-conjugate and the theorem is proved.

Concluding Remarks

The three theorems proved in this paper may be generalized to prove that the system consisting of a linear space X and a total subspace L of the space of all linear functionals defined on X is characterized to within isomorphism by its

lattice of L -closed subspaces, by the ring of L -continuous linear transformations of X into itself and, if the system L, X be identified with the system X, L , by its group of automorphisms. X_1, L_1 and X_2, L_2 are said to be isomorphic if there exist one-to-one linear transformations of all of X_1 into all of X_2 such that whenever x in X_1 corresponds to x' in X_2 and f in L_1 corresponds to f' in L_2 then $f(x) = f'(x')$. An L continuous linear transformation is a linear transformation T such that $F(x) = f(T(x))$ for all x in X is in L whenever f is. An L closed subspace is an intersection of null spaces of linear functionals in L . If we let L be the set of continuous functionals for a norm, we get as special cases the theorems of this paper. In the author's thesis, now in progress, the systems X, L are investigated systematically.

Due principally to the fact that distinct convex topologies may have the same set of continuous linear functionals, the isomorphism theorems cannot be extended, as they stand, to arbitrary convex linear topological spaces. The author expects to discuss the situation in detail in a later paper.

Eidelheit [2] shows that, at least in the case of complete spaces, any isomorphism between the rings R_1 and R_2 can be represented in the form $T' = UTU^{-1}$ where T is in R_1 , T' is in R_2 and U is a one-to-one bicontinuous linear transformation from all of X_1 into all of X_2 . The author expects to discuss the extent to which this is true in the group situation at a later date.

Other questions suggesting themselves for investigation include the following: (1) What properties must an abstract group (ring, lattice) have in order to be the group (ring, lattice) associated with some normed linear space or more generally some system X, L in accordance with the above theorems? (2) What special properties do the groups (rings, lattices) of complete, reflexive and other special kinds of normed linear space have? (3) Can we give simple characterizations of various well known normed linear spaces in terms of properties of the algebraic systems associated with them? (4) Do any of the three theorems imply any of the others without the intervention of Lemmas A and B? The derivation of the ring theorem from the group theorem looks as if it might be fairly easy.

HARVARD UNIVERSITY

BIBLIOGRAPHY

1. BANACH, S., *Théorie des Opérations Linéaires*, 1932.
2. EIDELHEIT, M., *On isomorphisms of rings of linear operators*, *Studia Mathematica*, vol. 9 (1940), pp. 97-105.
3. VEBLEN, O. AND YOUNG, J., *Projective Geometry*.
4. FICHTENHOLZ, G., *Sur les fonctionnelles linéaires continues au sens généralisé*, *Mathematicheskii Sbornik*, N. S., vol. 4 (1938), pp. 193-213.
5. HAUSDORFF, F., *Zur Theorie der linearen metrischen Räume*, *Journ. f. reine u. angew. Math.*, vol. 167 (1932) pp. 294-311.
6. SOBCZYK, A., *Projections in Minkowski and Banach spaces*, *Duke Mathematical Journal*, vol. 8 (1941), pp. 78-106.

NEUAUFBAU DER ENDETHEORIE

VON HANS FREUDENTHAL

(Received September 2, 1941)

Kompaktifizierungen topologischer Räume kennt man bereits aus anderen topologische Gebieten; z.B. ist das Bedürfnis an einem kompakten Substrat der geometrischen Untersuchungen wohl eine der Ursachen für die Ergänzung der euklidischen zur projektiven oder zur funktionentheoretischen Ebene. Man sieht an diesem Beispiel, daß die Kompaktifizierung sehr verschieden ausfallen kann. Topologisch wird man die funktionentheoretische Kompaktifizierung als die natürlichere ansehen; es liegt ja topologisch nicht der mindeste Grund vor, warum man gewisse divergente Folgen gegen den einen und gewisse gegen den andern unendlich fernen Punkt konvergieren lassen soll. Bei der unendlichen Geraden hingegen wird man topologisch gerade die Kompaktifizierung durch zwei unendlich ferne Punkte ("Endpunkte") vor der durch *einen* bevorzugen, da gar nicht einzusehen ist, warum man die nach verschiedenen Richtungen divergierenden Punktfolgen zusammenwerfen sollte. Ebenso wird man den unendlichen Zylinder durch zwei "Endpunkte" kompaktifizieren, die dreimal gelochte Sphäre (die Badehose) durch drei Endpunkte usw.

Von einer "natürlichen" Kompaktifizierung wird man also verlangen:

1. Das Unendlichferne, die Endpunktmenge, soll möglichst dünn sein (genauer: nulldimensional).

2. Das Unendlichferne soll möglichst weitgehend aufgespalten sein (ohne daß dabei gewisse allgemeine Raum-Axiome verletzt werden).

In meiner Dissertation¹ habe ich (zum Zweck topologisch-gruppentheoretischer Untersuchungen) zuerst dies Kompaktifizierungsproblem behandelt und gelöst (durch Einführung der "Endpunkte")² für die topologischen Räume R , die folgenden Bedingungen genügen:

- a) zweites Abzählbarkeitsaxiom,
- b) Kompaktheit im Kleinen,
- c) Zusammenhang im Kleinen,
- d) Zusammenhang.

In Fußnote 15 meiner Dissertation hatte ich bereits angekündigt, daß man die Bedingung c fallen lassen kann.

Herr L. Zippin hat³ andererseits gerade die Bedingung b abgeschwächt und ersetzt durch

- b') Semikompaktheit,⁴

¹ Berlin 1931: *Über die Enden topologischer Räume und Gruppen*, Math. Zeitschrift **33** (1931), 692-713.

² "Endpunkte" sind beiläufig in sehr speziellen Fällen von B. v. Kerékjártó (*Vorlesungen über Topologie*, Berlin 1923, S. 164) und von L. Zippin (Transactions Amer. Math. Soc. **31** (1929), 744-770, besonders 763) eingeführt worden.

³ *On semicompact spaces* (Amer. Journ. of Math. **57** (1935), 327-341). L. Zippin war bei seinen Untersuchungen teilweise unabhängig von den in Fußnote 1 zitierten.

⁴ In Wirklichkeit verlangt Zippin noch Metrisierbarkeit, was aber überflüssig ist (siehe 4.2 und 5.4).

d.h. die Existenz beliebig kleiner Umgebungen (zu jedem Punkte) mit kompakter Berandung. b' ist in der Tat die "wahre" Bedingung, wenn man wünscht, daß das Unendliche nulldimensional, also insbesondere jeder endliche Punkt in beliebig kleinen Umgebungen enthalten sein soll, die keine „unendlichfernen“ Punkte auf ihrem Rande besitzen.

In der vorliegenden Arbeit will ich das Versprechen aus meiner Dissertation erfüllen und die Theorie der Endpunkte aufbauen ohne die Bedingung c . Dabei werden ganz merkwürdige Schwierigkeiten auftauchen, die ich in 6.4–5 überwinden werde; dort liegt also der Schwerpunkt der Arbeit.

Es zeigt sich nun, daß man auch d fast ganz fallen lassen kann; man ersetzt d durch

$d')$ Kompaktheit des Komponentenraumes von R ,

d.h.: jede abnehmende Folge nichtleerer offener abgeschlossener Teilmengen von R soll einen nichtleeren Durchschnitt besitzen.

Der Bedingung d' genügen alle zusammenhängenden Räume und alle Räume mit nur endlich vielen Komponenten. Dagegen schließt d' aus, daß R aus unendlich vielen isolierten Komponenten *besteht*, d' schließt allerdings nicht notwendig aus, daß R unendlich viel Komponenten *enthält*, die sich nirgends häufen. Beispielsweise ist zulässig folgender Raum (Teilmenge der cartesischen Ebene), der zusammengesetzt ist aus den Komponenten

$$C_n : x = \frac{1}{n} \quad (n \text{ natürlich}),$$

$$D_n : x = 0, n < y < n + 1 \quad (n \text{ ganz});$$

hier häufen sich nämlich die D_n nirgends, und doch gilt d' , da jede offene abgeschlossene Menge, die *eine* D_n enthält, fast alle C_n , also auch *alle* D_n enthält.

Die Bedingung d' ist nun in der Tat unvermeidlich, wenn man einen Raum durch Endpunkte kompaktifizieren will. Nimmt man als R einen aus unendlich vielen isolierten Punkten bestehenden Raum, so bemerkt man sofort, daß die Aufspaltung des Unendlichen hier transfinit fortgesetzt werden muß, und daß man als Abschließung bestenfalls ein pathologisches (nicht dem 1. Abzählbarkeitsaxiom genügendes) Gebilde erhält.

Auch die Zippinsche Abschwächung b' übernehme ich. Wir sahen bereits, daß auch hier keine weitere Abschwächung möglich ist, wenn die Endpunktmenge nulldimensional sein soll.

Meine Voraussetzungen lauten also:

- a) zweites Abzählbarkeitsaxiom,
- b') Semikompaktheit,
- d') Kompaktheit des Komponentenraumes.

Man kann noch das zweite Abzählbarkeitsaxiom fallen lassen, wenn man die Semikompaktheit durch die Semibikompaktheit ersetzt und kein *kompaktes*, sondern ein *bikompaktes* Resultat anstrebt. Sehr wichtig ist das nicht; Zufügung der Endpunkte macht den Raum nämlich auch dann bikompakt, wenn d'

garnicht gilt. Die Hauptschwierigkeit beim Beweise (siehe 6.4–5) besteht gerade darin, zu zeigen, daß aus einem Raum mit zweitem Abzählbarkeitsaxiom, bei Gültigkeit von d' , wieder ein Raum mit zweitem Abzählbarkeitsaxiom entsteht (der dann als bikompakter Raum kompakt sein muß). Trotzdem habe ich mich soweit möglich auf den etwas allgemeineren bikompakten Standpunkt gestellt.

Dadurch daß ich mich auf die unumgänglichen Forderungen a , b' , d' habe beschränken können, habe ich, wie mir scheint, die Endentheorie zu einem gewissen Abschluß gebracht. Meine Methode weicht anfangs nicht nennenswert von der meiner Dissertation ab (erst in 6.4–5 kommt der eigentliche Unterschied). Die "Endpunkte" werden natürlich wieder durch absteigende Folgen von offenen Mengen mit kompakter Berandung definiert.⁵ Man muß diese Folgen natürlich gewissen Feinheitbeschränkungen unterwerfen. In meiner Dissertation und bei Zippin geschieht das auf Grund der Forderung c . Der Zusammenhang im Kleinen bewirkt nämlich, daß man sich auf *Gebiete* für die erzeugenden offenen Mengen beschränken kann, und daß diese Gebiete bei weiterem Absteigen in wesentlich nur endlich viel Teilgebiete zerfallen. So fließt aus dem Zusammenhang im Kleinen einerseits der atomistische Charakter der verwendeten Folgen, andererseits die Kompaktheit des Resultats.

Steht einem c nicht zur Verfügung, so kann man den atomistischen Charakter durch Maximalitätsforderungen erzwingen. Es schien mir aber zweckmäßiger und anschaulicher, von den erzeugenden Folgen $G_1 \supset G_2 \supset \dots$ zu verlangen, daß für je zwei offene Mengen O, P mit kompakter Berandung und mit $\bar{O} \subset P$ gilt:

ist einmal $O \cap G_n \neq \emptyset$, so ist $G_n \subset P$ für fast alle n .

In meiner Dissertation habe ich auch ein Charakterisierungsproblem behandelt. Ist R^* ein Kompaktum und R überall dicht in R^* , so kann man sich fragen, ob vielleicht R^* gerade die Kompaktifizierung von R durch die Endpunkte ist. Die notwendigen und hinreichenden Bedingungen meiner Dissertation lauten:

α) $R^* \setminus R$ ist abgeschlossen und nulldimensional,

β) eine Gebiets-Umgebung eines Punktes von $R^* \setminus R$ wird durch $R^* \setminus R$ nicht zerlegt.

Zippin hat α a.a.O. abgeschwächt zu

$\alpha')$ $R^* \setminus R$ ist ein total zusammenhangsloses F_σ .

In der vorliegenden Arbeit lauten die Charakterisierungsbedingungen:

$\alpha')$ $R^* \setminus R$ ist nulldimensional,

$\beta')$ die Umgebungen eines Punktes p von $R^* \setminus R$ werden durch $R^* \setminus R$ nicht zerlegt in zwei in R offene Mengen, die beide p als Häufungspunkt besitzen.

Veranlaßt wurde die vorliegende Arbeit durch schöne Untersuchungen des

⁵ Natürlich muß man sich auf offene Mengen mit kompakter Berandung beschränken, wenn man nicht zu pathologischen Resultaten gelangen will. Das ist verschiedentlich übersehen worden.

Herrn J. de Groot, der sich mit folgendem Problem beschäftigt hat: Seien A und A' homöomorph, $A \cap B = A' \cap B' = \circ$. Wann läßt sich *jede* homöomorphe Beziehung zwischen A und A' erweitern zu einer homöomorphen Beziehung zwischen $A \cup B$ und $A' \cup B'$?

Verlangt man, wie Herr de Groot es in der Tat tut, daß $A \cup B$ und $A' \cup B'$ Kompakta seien, und unterwirft man die Mengen weiteren Bedingungen, die B bzw. B' gerade als Endpunktmenge von A bzw. A' charakterisieren, so erzwingt man in der Tat die verlangte Fortsetzbarkeit jeder Homöomorphie.

Herr de Groot kam unmittelbar von seiner Fragestellung aus, ohne Kenntnis meiner und der Zippinschen Untersuchungen, zu Resultaten die sich mit den Zippinschen überschneiden. In seiner Note⁶ verlangt er (außer der Überall-dichtheit):

$\alpha')$ B und B' sind nulldimensional,

$\beta')$ B bzw. B' zerlegen $A \cup B$ bzw. $A' \cup B'$ nicht im Kleinen. In weiteren, noch unpublizierten Untersuchungen schwächt er β'' weiter ab. Seine neuen Bedingungen sind gerade mit den früher genannten α', β' identisch.

Ich bemerke aber, daß Herrn de Groot für diesen Satz (und damit auch für den früher genannten Charakterisierungssatz) in vollem Umfang die Priorität zukommt, und daß es erst diese sehr allgemeinen Resultate des Herrn de Groot waren, die mich veranlaßten, die Untersuchungen meiner Dissertation wiederaufzunehmen. Anfangs schien es mir zwar, als ob die größere Allgemeinheit der Resultate des Herrn de Groot durch die Andersartigkeit der Fragestellung bedingt war; dann gelang es aber, den de Groot'schen Fortsetzungssatz als Charakterisierungssatz in eine verallgemeinerte Endentheorie einzubauen.

Wir wenden zum Schluß (in 8) die Begriffe an auf die gruppentheoretische Frage, die den Ausgangspunkt meiner Dissertation bildete. In meiner Dissertation habe ich nämlich gezeigt, daß ein Gruppenraum, der a-d genügt, höchstens zwei Endpunkte besitzen kann. Zippin hat diesen Satz ausgedehnt auf Gruppenräume, die a, b, c', d genügen; er mußte meine Überlegungen dabei ziemlich abändern. Durch erneute, sehr weitgehende Abänderung und Vertiefung der Methode kann ich nun beweisen, daß ein Gruppenraum, der

- 1) dem zweiten Abzählbarkeitsaxiom genügt,
- 2) semikompakt und
- 3) zusammenhängend

ist, höchstens zwei Endpunkte besitzt. Da R kompakt ist und $R^* \setminus R$ aus höchstens zwei Punkten besteht, ist R notwendig im Kleinen kompakt. Daraus folgt der merkwürdige Satz:

Ein Gruppenraum, der 1-3 genügt, ist von selber im Kleinen kompakt.

Bezeichnungen

\circ = leere Menge.

$A \cup B$ = Vereinigung von A und B .

⁶ Proc. Akad. Amsterdam 44 (1941), ——— = Indagationes Math. 3 (1941), -

$A \cap B$ = Durchschnitt von A und B .

\cup = Zeichen für die Vereinigungsbildung über eine Menge von Mengen.

\cap = Zeichen für die Durchschnittsbildung über eine Menge von Mengen.

$a \in A$: a ist Element von A .

$a \notin A$: a ist nicht Element von A .

$A \setminus B$: Gesamtheit der $a \in A$, $a \notin B$.

\bar{A} = abgeschlossene Hülle von A .

$R(M)$ = Rand von M .

Gotische Buchstaben verwenden wir häufig für Mengen, deren Elemente Mengen sind.

1. Allgemeine topologische Begriffe

1.1. R sei im Folgenden ein topologischer Raum, d.h. in R sei ein System von *offenen* Mengen gegeben, zu dem \emptyset , R , mit zwei Mengen ihr Durchschnitt, mit beliebig vielen ihre Vereinigung gehört. Die Komplemente der offenen Mengen heißen *abgeschlossen*. *Umgebung* einer Menge heißt jede sie enthaltende offene Menge. Der Durchschnitt aller M enthaltenden abgeschlossenen Mengen heißt \bar{M} , die *abgeschlossene Hülle* von M . Ein Punkt gehört zum Rand von M , $R(M)$, dann und nur dann, wenn jede seiner Umgebungen Punkte von M und Punkte von $R \setminus M$ enthält.

1.2. Wir verwenden wiederholt folgende einfache Folgerungen: Sind zwei offene Mengen zueinander fremd, so ist jede zur abgeschlossenen Hülle der anderen fremd. Ist O offen, so gilt $R(O) = \bar{O} \setminus O$.

1.3. Wir setzen im Folgenden stets die Trennungseigenschaft \mathfrak{T} voraus: Je zwei Punkte lassen sich durch fremde Umgebungen trennen.—Insbesondere ist dann jede abgeschlossene Menge der Durchschnitt ihrer Umgebungen, und ist jede einpunktige Menge abgeschlossen.

1.4. *Basis* von R heißt ein System, aus dem sich alle offenen Mengen durch Vereinigungsbildung erzeugen lassen. Das zweite Abzählbarkeitsaxiom lautet: R besitzt eine abzählbare Basis. Ein solcher Raum heißt auch separabel. *Umgebungsbasis* von M heißt ein System \mathfrak{B} von Umgebungen von M , wenn zu jeder Umgebung U von M ein $V \in \mathfrak{B}$ mit $V \subset U$ existiert.

1.5. R heißt *regulär*, wenn zu jedem Punkte a und zu jeder Umgebung U von a eine Umgebung V von a existiert mit $\bar{U} \subset V$. R heißt *normal*, wenn das Entsprechende für jede abgeschlossene Menge gilt. Reguläre separable Räume sind bekanntlich normal.

1.6. Ein System \mathfrak{a} von Mengen aus R heißt *endlich gebunden*, wenn der Durchschnitt von je endlich viel Mengen aus \mathfrak{a} nicht leer ist. \mathfrak{a} heißt *gebunden*, wenn ganz \mathfrak{a} einen nicht leeren Durchschnitt besitzt.

R heißt *bikompakt*, wenn jedes endlich gebundene System von abgeschlossenen Mengen aus R auch gebunden ist. Bikompaktheit ist bekanntlich äquivalent mit der *Überdeckungseigenschaft*: Jede Überdeckung von R mit offenen Mengen enthält eine endliche Überdeckung.

Ein separabler bikompakter Raum heißt ein *Kompaktum*.

1.7. R heißt *semibikompakt* bzw. *Semikompaktum*, wenn eine Basis K von R existiert, derart daß jede Menge aus K eine bikompakte Menge bzw. ein Kompaktum als Rand besitzt.

1.8. Eine Menge, die offen und abgeschlossen zugleich ist, heißt ein *Brocken*. Das Komplement eines Brockens ist wieder ein Brocken. Ein Raum R heißt *zusammenhängend*, wenn er keine Brocken außer \circ und R enthält. *Komponente* von R heißt jede maximale zusammenhängende Teilmenge. Ein Raum heißt *nulldimensional*, wenn seine Brocken eine Basis bilden.

1.9. Ein Kompaktum kann nur abzählbar viel Brocken besitzen. Denn das zweite Abzählbarkeitsaxiom liefert eine abzählbare Basis für die Brocken; wegen des Überdeckungssatzes läßt sich aber *jeder* Brocken bereits aus *endlich* vielen dieser Basisbrocken erzeugen.

1.10. M heißt überall dicht in R , wenn jede offene Menge von R Punkte von M enthält.

2. Die erzeugenden Systeme

2.1. Wir betrachten topologische Räume R , in denen eine (zunächst undefinierte) Relation zwischen offenen Mengen gegeben ist, die wir mit \subset bezeichnen. Die Beziehung \subset genügt den Bedingungen:

1. Ist $O \subset P$, so ist $\bar{O} \subset P$.
2. Ist $O_1 \subset P$, $O_2 \subset P$, so ist $O_1 \cup O_2 \subset P$.
3. Ist $O \subset P_1$, $O \subset P_2$, so ist $O \subset P_1 \cap P_2$.
4. Ist $O \subset P$, so ist $R \setminus \bar{P} \subset R \setminus \bar{O}$.
5. Ist $O_1 \subset O_2$, $O_2 \subset P_2$, $P_2 \subset P_1$, so ist $O_1 \subset P_1$.

Ein System von Relationen \subset , das diesen 5 Bedingungen genügt heißt ein \mathfrak{D} -System (auch $\mathfrak{D}(R)$). Gilt 2.1.5 nicht notwendig, so heißt es ein \mathfrak{D}' -System (auch $\mathfrak{D}'(R)$).

Aus 2.15 folgt, daß jedes nichtleere \mathfrak{D} -System die Relationen $\circ \subset O \subset R$ enthält.

Einen mit einem \mathfrak{D} -System versehenen Raum nennen wir einen \mathfrak{D} -Raum. Man kann jeden Raum R zu einem \mathfrak{D} -Raum machen, wenn man $O \subset P$ dann und nur dann vorschreibt, falls $\bar{O} \subset P$ (man sieht leicht, daß allen Forderungen genügt ist). Man ist aber dazu nicht verpflichtet; man kann als \mathfrak{D} -System von R auch eine echte Teilmenge davon nehmen.

Ein \mathfrak{D} - bzw. \mathfrak{D}' -System von R erzeugt in jedem Teilraum S wieder ein \mathfrak{D} - bzw. \mathfrak{D}' -System. Man setze nämlich $O \cap S \subset P \cap S$ in S dann und nur dann, wenn $O \subset P$ in R gilt. Beim Übergang zu Teilräumen legen wir stets das induzierte System zugrunde.

Man kann \subset anschaulich auch als einen *qualitativen* Abstandsbegriff deuten; man deute nämlich $O \subset P$ als: O und $R \setminus \bar{P}$ besitzen einen *Abstand*.

2.2. Sei in R ein \mathfrak{D}' -System gegeben. Man konstruiere folgendes System \mathfrak{D} : Dann und nur dann ist die Relation $O \subset P$ in \mathfrak{D} , wenn es in \mathfrak{D}' eine Relation $O_1 \subset P_1$ gibt mit $O \subset O_1 \subset P_1 \subset P$. Wie man leicht sieht, ist das neue System in der Tat ein \mathfrak{D} -System. \mathfrak{D}' heißt dann *Basis* von \mathfrak{D} und \mathfrak{D} heißt von \mathfrak{D}' erzeugt.

2.3. Eine Menge g von offenen Mengen des \mathfrak{D} -Raumes R heit eine *Erzeugende*, wenn gilt:

1. g ist endlich gebunden.

2. Zu jedem $G \in g$ existiert ein $G' \in g$ mit $G' \subset G$.

3. Zu je zwei offenen Mengen O, P mit $O \subset P$ und mit $O \cap G \neq \circ$ fr alle $G \in g$ existiert ein $G_0 \in g$ mit $G_0 \subset P$.

2.4. Ist g Erzeugende und $G_1 \in g, G_2 \in g$, so existiert $G_3 \in g$ mit $G_3 \subset G_1 \cap G_2$.

BEWEIS: Nach 2.3. existieren $G'_1, G'_2 \in g$ mit $G'_i \subset G_i$. Wir setzen $O = G'_1 \cap G'_2$, $P = G_1 \cap G_2$. Nach 2.1.3 und 2.1.5 ist $O \subset P$ und nach 2.3.1 ist $O \cap G \neq \circ$ fr alle $G \in g$. Nach 2.3.3 existiert also ein $G_3 \in g$ mit $G_3 \subset P = G_1 \cap G_2$, w.z.b.w.

2.5. Wir definieren:

$g < O$: $G \subset O$ fr ein gewisses $G \in g$.

$h < g$: $h < G$ fr alle $G \in g$.

$gh \neq \circ$: $G \cap H \neq \circ$ fr alle $G \in g$ und $H \in h$.

2.6. Aus $g < h$ folgt $gh \neq \circ$. Aus $gh \neq \circ$ folgt $g < h$ (also auch $h < g$).

BEWEIS: Erste Hlfte klar.—Sei nun $gh \neq \circ$. Sei $H \in h$. Wir bestimmen nach 2.3.2 ein $H' \in h$ mit $H' \subset H$. Nach Voraussetzung ist $G \cap H' \neq \circ$ fr alle $G \in g$. Also existiert nach 2.3.3 ein $G \in g$ mit $G \subset H$. Also $g < H$. Das gilt fr alle $H \in h$. Also $g < h$.

2.7. Schreibt man $g = h$ statt $gh \neq \circ$ (oder $g < h$), so ersieht man aus 2.6, da die blichen Rechengesetze fr das Gleichheitszeichen gelten.

3. Der Raum R^*

3.1. Wir beabsichtigen, die Erzeugenden von R (unter Bercksichtigung der definierten Gleichheit) als Punkte eines neuen Raumes R^* zu deuten. Zunchst definieren wir:

$O^* =$ Menge aller $g < O$. Speziell $R^* =$ Menge aller Erzeugenden von R .

$O^* \subset P^*$ dann und nur dann, wenn $O \subset P$.

3.2.1. Aus $O \subset P$ folgt $O^* \subset P^*$.

2. $(O \cap P)^* = O^* \cap P^*$.

3. $(O \cup P)^* \supset O^* \cup P^*$.

BEWEIS: 3.2.1 ist klar. 3.2.2: Sei $g \in (O \cap P)^*$. Dann $g < O \cap P$, also $g < O, g < P$. Also $g \in O^*, g \in P^*$, also $g \in O^* \cap P^*$.—Sei umgekehrt $g \in O^* \cap P^*$. Dann $g \in O^*, g \in P^*$. Also $g < O, g < P$. Also existiert $G_1 \in g, G_1 \subset O; G_2 \in g, G_2 \subset P$; also auch nach 2.4 ein $G_3 \in g$ mit $G_3 \subset G_1 \cap G_2$. Dann ist aber $G_3 \subset O \cap P$, also $g < O \cap P$, also $g \in (O \cap P)^*$, w.z.b.w.—Nun 3.2.3: Sei $g \in O^* \cap P^*$. Dann $g < O$ oder $g < P$, also $g < O \cup P$, also $g \in (O \cup P)^*$.

3.3. R^* wird zu einem topologischen Raum durch die Festsetzung: Die O^* bilden eine Basis von R^* . (Folgt aus 3.2.2.) Die Randbildung in R^* heie \mathfrak{R}^* .

3.4.1. Die G^* mit $G < g$ bilden eine Umgebungsbasis von g (siehe 1.4). (Klar.)

2. Man kann jedes g erzeugen unter ausschlielicher Verwendung solcher G , die in den Relationen der Basis \mathfrak{D}' von \mathfrak{D} auftreten. (Man ersetze nur je zwei Mengen $G, H \in g$ mit $G \subset H$ durch G_1 und H_1 , wo $G_1 \subset H_1$ zu \mathfrak{D}' gehren und

$G \subset G_1 \subset H_1 \subset H$ gelten möge.) 3. Die O^* , die in den Relationen von \mathfrak{D}' auftreten, bilden eine Basis von R^* . (Folgt aus 3.4.1–2.)

3.5. $g \in \bar{O}^*$ dann und nur dann, wenn $O \cap G \neq \circ$ für alle $G \in g$.

BEWEIS: Nach 1.2 und 3.4.1 ist $g \in \bar{O}^*$ äquivalent mit: $G^* \cap O^* \neq \circ$ für alle $G \in g$. Das ist aber nach 3.4.2 äquivalent mit: $G \cap O \neq \circ$ für alle $G \in g$, w.z.b.w.

3.6. Aus $\bar{O} \subset P$ folgt $\bar{O}^* \subset P^*$.

BEWEIS: Sei $g \in \bar{O}^*$. Dann nach 3.5: $G \cap O \neq \circ$ für alle $G \in g$. Also nach 2.3.3: es existiert ein $G_0 \in g$, $G_0 \subset P$. Also $g \subset P$, $g \in P^*$, w.z.b.w.

3.7. $(R \setminus \bar{O})^* = R^* \setminus \bar{O}^*$.

BEWEIS:

$$g \in R^* \setminus \bar{O}^*$$

ist äquivalent mit

$$g \notin \bar{O}$$

oder nach 3.5 mit der

$$\text{Existenz eines } G \in g \text{ mit } G \cap O = \circ,$$

und die ist äquivalent mit der

$$\text{Existenz eines } G \in g \text{ mit } G \subset R \setminus \bar{O},$$

und die ist äquivalent mit

$$g \in (R \setminus \bar{O})^*, \text{ w.z.b.w.}$$

3.8. R^* ist regulär (erfüllt also auch T).

BEWEIS: Sei O^* Umgebung von g . Nach 3.4 und 3.2.1 ist $G \subset O$ für ein gewisses $G \in g$. Nach 2.3.2 existiert $G' \in g$ mit $G' \subset G$. Nach 2.1.1 ist $\bar{G}' \subset G$ und nach 3.6: $\bar{G}'^* \subset G^*$. Also auch $\bar{G}'^* \subset O^*$ und \bar{G}'^* ist die Umgebung von g , die die Regularität gewährleistet.

4. Die Bikompaktheit von R^*

4.1. R heißt \mathfrak{D} -regulär, wenn zu jedem Punkte a und jeder Umgebung V von a eine Umgebung U von a existiert mit $U \subset V$.

R heißt \mathfrak{D} -normal, wenn zu jeder Relation $O \subset P$ eine Relation $O \subset Q \subset P$ existiert.

R heißt \mathfrak{D} -separabel, wenn $\mathfrak{D}(R)$ eine abzählbare Basis $\mathfrak{D}'(R)$ besitzt.

4.2. Ein \mathfrak{D} -reguläres R ist auch regulär. Ein \mathfrak{D} -reguläres und \mathfrak{D} -separables R ist auch separabel und normal.

BEWEIS: Erster Teil folgt aus 2.1.1.—Zweiter Teil: Sei $\mathfrak{D}'(R)$ die abzählbare Basis von $\mathfrak{D}(R)$. Sei \mathfrak{P} die (abzählbare) Menge aller P , die in Relationen $O \subset P$ aus \mathfrak{D}' auftreten. Sei Q irgendeine offene Menge. Wegen der \mathfrak{D} -Regularität existiert zu jedem $a \in Q$ eine Umgebung U von a , $U \subset Q$. Gemäß dem Begriff der Basis existiert in \mathfrak{D}' eine Relation $O \subset P$ mit $U \subset O \subset P \subset Q$. Also existiert sicher ein $P \in \mathfrak{P}$ mit $a \in P \subset Q$. Die Vereinigung aller solcher P

(über alle $a \in Q$) liefert Q . Also läßt sich jede offene Menge Q als Vereinigung von Mengen von \mathfrak{B} darstellen. Da \mathfrak{B} abzählbar war, ist R separabel. Aus der Separabilität und Regularität folgt die Normalität.

4.3. $g(a)$ sei die Menge der Umgebungen von a .

4.4. R sei \mathfrak{D} -regulär. Dann ist g eine topologische Abbildung von R in R^* . Dabei ist O die Urbildmenge von O^* .

BEWEIS: 1. g ist eine Erzeugende: 2.3.1–2 sind evident wegen der \mathfrak{D} -Normalität. Mögen O und P den Voraussetzungen von 2.3.3 genügen; also $O \subset P$ und $O \cap G \neq \circ$ für alle $G \in \mathfrak{g}$. Wäre $a \notin P$, so wäre nach 2.1.1 auch $a \notin \bar{O}$, also $a \in R \setminus \bar{O}$, also $R \setminus \bar{O} \in g(a)$, aber $(R \setminus \bar{O}) \cap O = \circ$ im Widerspruch zur Voraussetzung über O . Also notwendig $a \in P$, also $P \in g(a)$, und P ist brauchbar als das in 2.3.3 verlangte G_0 .

2. Sei $a \neq b$. Dann gibt es wegen \mathfrak{T} : $U \in g(a)$, $V \in g(b)$, $U \cap V = \circ$. Also $g(a) \neq g(b)$. Also ist g eineindeutig.

3. Wir bestimmen die Urbildmenge von O^* , d.h. die Menge aller a mit $g(a) \in O^*$ oder $g(a) < O$. Das sind aber nach Definition gerade die a aus O . Also ist O das Urbild von O^* . Das liefert auch die Stetigkeit von g . Auch die Umkehrung von g ist stetig, denn $g(O) = O^* \cap g(R)$ ist offen in $g(R)$.

4.5. Wir können und wollen R auch als Teilmenge von R^* deuten. Wir haben dann auch $O^* \cap R = O$.

4.6. In einem \mathfrak{D} -regulären R gilt: (Ergänzung zu 3.2.1.) Aus $O^* \subset P^*$ folgt $O \subset P$. (Folgt aus 4.5.)

4.7. R sei \mathfrak{D} -regulär. R^* wird zu einem \mathfrak{D} -Raum durch die Festsetzung: Die Relationen $O^* \subset P^*$ (mit $O \subset P$) bilden eine Basis $\mathfrak{D}'(R^*)$ von R^* .

BEWEIS: Es genügt, zu zeigen, daß für $\mathfrak{D}'(R^*)$ die Bedingungen 2.1.1–4 erfüllt sind:

(2.1.1:) Sei $O^* \subset P^*$. Nach Definition ist dann $O \subset P$, also nach 2.1.1 (für $\mathfrak{D}(R)$) $\bar{O} \subset P$, also nach 3.6: $\bar{O}^* \subset P^*$, w.z.b.w.

(2.1.2:) Seien $O_1^* \subset P^*$, $O_2^* \subset P^*$. Definitionsgemäß ist $O_1 \subset P$, $O_2 \subset P$, also nach 2.1.2 (für $\mathfrak{D}(R)$) $O_1 \cup O_2 \subset P$. Nach 3.2.3 ist $O_1^* \cup O_2^* \subset (O_1 \cup O_2)^*$, also auch $\subset P^*$, w.z.b.w.

(2.1.3:) Analog unter Verwendung von 3.2.2.

(2.1.4:) Sei $O^* \subset P^*$, also $O \subset P$. Nach 2.1.4 (für $\mathfrak{D}(R)$) ist $R \setminus \bar{P} \subset R \setminus \bar{O}$, also $(R \setminus \bar{P})^* \subset (R \setminus \bar{O})^*$ und nach 3.7: $R^* \setminus \bar{P}^* \subset R^* \setminus \bar{O}^*$, w.z.b.w.

4.8. R sei \mathfrak{D} -regulär. Dann läßt sich R auch hinsichtlich des \mathfrak{D} -Systems als Teilraum von R^* auffassen. (Klar.)

4.9. SATZ I: R sei \mathfrak{D} -regulär. Dann ist auch R^* \mathfrak{D} -regulär. Ist R oben-
drein \mathfrak{D} -separabel, so ist auch R^* separabel und \mathfrak{D} -separabel. R^* ist eine Fort-
setzung von R , und R ist überall dicht in R^* .

(Folgt aus 3.4.3, 3.8, 4.2 und 4.8.)

4.10. R sei \mathfrak{D} -regulär und \mathfrak{D} -normal. Sei A abgeschlossen in R^* , $g \notin A$. Dann gibt es eine Folge U_n^* von Umgebungen von A mit $U_1 \supset U_2 \supset \dots$, derart daß $g \notin U_1^*$.

BEWEIS: Wir wählen $V \in \mathfrak{g}$ mit $\bar{V}^* \cap A = \circ$ und $W_1 \in \mathfrak{g}$ mit $W_1 \subset V$. Auf

grund der \mathfrak{D} -Normalität bestimmen wir induktiv ein W_{n+1} mit $W_n \subset W_{n+1} \subset V$. Wir setzen $U_n = R \setminus \overline{W_n}$. Dann liefert 2.1.4: $U_1 \supset U_2 \supset \dots \supset R \setminus V$. Weiter liefert 3.2.1: $U_n^* \supset (R \setminus V)^* \supset A$. Also sind die U_n in der Tat Umgebungen von A . Wegen $g < V$ ist $g \notin U_1^*$. Also erfüllen die U_n alle Wünsche.

4.11. SATZ II: *R sei \mathfrak{D} -regulär und \mathfrak{D} -normal. Dann ist R^* bikompakt. Ist R obendrein \mathfrak{D} -separabel, so ist R^* ein Kompaktum.*

BEWEIS: α sei eine endlich gebundene Menge abgeschlossener Mengen von R^* . Wir beweisen die Gebundenheit von α .

Sei $A \in \alpha$. Zu jedem $g \in A$ bestimmen wir eine Folge $U_n^{g,A}$ gemäß 4.10. u sei die Menge aller dieser offenen Mengen, wo A ganz α und g ganz $R \setminus A$ durchläuft. Wir zeigen, daß u die Eigenschaften 2.3.1–2 eines g besitzt:

(2.3.1:) $U_{n_k}^{g,A_{\kappa}}$ ($\kappa = 1, \dots, k$) seien endlich viel Mengen aus u . Es gibt ein $\mathfrak{h} \in \cap_{\kappa} A_{\kappa}$. Dann auch $\mathfrak{h} \in (U_{n_k}^{g,A_{\kappa}})^*$. Also existiert ein $H \in \mathfrak{h}$ mit $H \subset U_{n_k}^{g,A_{\kappa}}$ ($\kappa = 1, \dots, k$). Da H nichtleer ist, ist auch $\cap_{\kappa} U_{n_k}^{g,A_{\kappa}}$ nichtleer. Also ist u endlich gebunden.

(2.3.2:) Sei $U_n^{g,A} \in u$. Dann ist $U_{n+1}^{g,A} \subset U_n^{g,A}$. Also gilt 2.3.2.

Wir bilden aus u alle Durchschnitte zu je endlich vielen; so entsteht \mathfrak{v} . Auch \mathfrak{v} erfüllt 2.3.1 (trivial) und 2.3.2 (wegen 2.3 und 2.5 folgt nämlich aus $O_{\kappa} \subset P_{\kappa}$: $\cap_{\kappa=1}^k O_{\kappa} \subset \cap_{\kappa=1}^k P_{\kappa}$).

Dagegen braucht 2.3.3 nicht zu gelten. Sei nun O, P ein Paar, das 2.3.3 in bezug auf \mathfrak{v} verletzt. Also jedenfalls $O \subset P$, $O \cap G \neq \emptyset$ für alle $G \in \mathfrak{v}$. Wir bestimmen auf grund der \mathfrak{D} -Normalität eine Folge $P_n: P_1 = P, O \subset P_{n+1} \subset P_n$ und adjungieren die P_n zu \mathfrak{v} . So entsteht ein System u' . u' erfüllt 2.3.1 (da \mathfrak{v} es erfüllte) und 2.3.2 (wegen der Wahl der P_n). Der wesentliche Unterschied zwischen u' und u ist aber der, daß das Paar O, P in bezug auf u' nicht mehr 2.3.3 verletzt.

Unter Verwendung einer Wohlordnung kann man u also zu einer Menge g ergänzen, die 2.3.1–3 erfüllt. $g \in R^*$. $g < \text{alle } U \text{ von } u$. Also $g \in \text{alle } U^*$ mit $U \in u$. Also $g \in \cap_{U \in u} U^*$. Wäre $g \in \cap_{A \in \alpha} A$, so gäbe es ein $A_0 \in \alpha$ mit $g \notin A_0$ und ein $U \in u$ mit $g \notin \overline{U^*}$, und das lieferte einen Widerspruch. Also ist $g \in \cap_{A \in \alpha} A$, und α ist gebunden, w.z.b.w.

Der Rest des Satzes folgt nun aus Satz I.

4.12. SATZ III: *In R^* gilt unter den Voraussetzungen von Satz II: Aus $\bar{O} \subset P$ folgt $O \subset P$.*

BEWEIS: Zu jedem $g \in \bar{O}$ gibt es ein $H \in g$, $H^* \subset P$; ferner ein $G \in g$, $G \subset H$. Endlich viele der zugehörigen G^* überdecken das bikompakte \bar{O} ; z.B. G_1^*, \dots, G_k^* . Wegen $G_{\kappa} \subset H_{\kappa}$ und 2.1.1–2 ist $G = \bigcup G_{\kappa} \subset \bigcup H_{\kappa} = H$. Also $G \subset H$ und $O \subset G^* \subset H^* \subset P$, also nach 2.5 und 4.7: $O \subset P$, w.z.b.w.

5. Die Endpunkte

5.1. Wir betrachten nun spezielle Systeme \mathfrak{D} .

Sei \mathfrak{R} das System aller offenen Mengen von R mit bikompaktem Rand und \mathfrak{B}

das aller mit *leerem* Rand, d.h. das System aller Brocken.—Aus $O \in \mathfrak{R}$ folgt $R \setminus \bar{O} \in \mathfrak{R}$.

$\mathfrak{D}'_{\mathfrak{R}}$ ist so definiert: Man setzt $O \subset P$, wenn $\bar{O} \subset P$ und $O \in \mathfrak{R}$, $P \in \mathfrak{R}$.

\mathfrak{D}'_3 ist so definiert: Man setzt $O \subset O$ für alle $O \in \mathfrak{Z}$.

Daß beidemal 2.1.1–4 gelten, ist leicht zu sehen.

\mathfrak{D}_3 resp. $\mathfrak{D}_{\mathfrak{R}}$ sind die von \mathfrak{D}'_3 resp. $\mathfrak{D}'_{\mathfrak{R}}$ erzeugten Systeme.

5.2. Sei R semibikompakt. Sei $O \in \mathfrak{R}$, P offen, $\bar{O} \subset P$. Dann gibt es ein $P_1 \in \mathfrak{R}$ mit $\bar{O} \subset P_1 \subset \bar{P}_1 \subset P$.

BEWEIS: Zu jedem Punkt von $\mathfrak{R}(O)$ gibt es eine \mathfrak{R} -Umgebung U mit $\bar{U} \subset P$. Endlich viele U_1, \dots, U_k , überdecken $\mathfrak{R}(O)$. $P = O_1 \cup U_1 \cup \dots \cup U_k$ erfüllt die Forderung.

5.3. Sei R semibikompakt. $O \subset P$ gemäß $\mathfrak{D}_{\mathfrak{R}}$ dann und nur dann, wenn ein $O_1 \in \mathfrak{R}$ existiert mit $O \subset O_1 \subset \bar{O}_1 \subset P$.

BEWEIS: Nur dann: Da $\mathfrak{D}'_{\mathfrak{R}}$ Basis von $\mathfrak{D}_{\mathfrak{R}}$ ist, gibt es $O_1, P_1 \in \mathfrak{R}$ mit $O \subset O_1 \subset P_1 \subset P$, also auch $O \subset O_1 \subset \bar{O}_1 \subset P$. Dann: Nach 5.2 existiert $P \in \mathfrak{R}$ mit $\bar{O}_1 \subset P_1 \subset P$. Also $O \subset O_1 \subset P_1 \subset P$, also $O \subset P$ nach 2.1.5.

5.4. Ein semibikompaktes R ist $\mathfrak{D}_{\mathfrak{R}}$ -normal.

BEWEIS: Es genügt für $O, P \in \mathfrak{R}$ aus $O \subset P$ die Existenz einer Relation $O \subset Q \subset P$ abzuleiten. Die ergibt sich aber aus 5.2.

5.5. Ein semibikompaktes R ist auch $\mathfrak{D}_{\mathfrak{R}}$ -regulär.

BEWEIS: Sei U Umgebung von a . Es gibt eine Umgebung V von a , $V \in K$. $(R \setminus \bar{V}) \subset R \setminus (a)$. Also gibt es nach 5.2 ein $P_1 \in \mathfrak{R}$ mit $(R \setminus \bar{V}) \subset P_1 \subset R \setminus (a)$. Dann ist $W = R \setminus \bar{P}_1$ Umgebung von a , $W \in K$ und $\bar{W} \subset U$, w.z.b.w.

5.6. Die Bezeichnung R^* reservieren wir nun für den aus R mittels $\mathfrak{D}_{\mathfrak{R}}$ erzeugten Raum (für den Fall, daß R semibikompakt ist). Wir nennen R^* auch die Abschließung von R durch seine Endpunkte. $R^* \setminus R$ heißt die Menge der Endpunkte.

Es ist klar, daß man bei der Erzeugung von R^* an die g die zusätzliche Forderung $g \subset \mathfrak{R}$ stellen darf. Das werden wir meistens auch stillschweigend tun.

Bei \mathfrak{D}_3 sind die Voraussetzungen von Satz I–II nicht mehr erfüllt (keine \mathfrak{D} -Regularität!). Trotzdem bleibt vieles erhalten. Wir nennen den aus R mittels \mathfrak{D}_3 erzeugten Raum $Z(R)$, den Komponentenraum von R . In der Tat sind in einfachen Fällen die Komponenten von R Elemente von $Z(R)$; allerdings zeigt das Beispiel der Einleitung, daß das nicht immer gilt: die Komponenten D_n bilden einen Punkt von $Z(R)$.

Für die \mathfrak{D}_3 -Erzeugenden sind die Forderungen 2.3.1–3 besonders leicht zu erfüllen; sie reduzieren sich darauf, daß g aus Brocken besteht und daß g endlich verbunden und maximal ist.

5.7. Die Abbildung $\zeta(O)$ ordne jeder Menge den kleinsten sie enthaltenden Brocken zu. Für $g \in R^*$ setzen wir $\zeta(g) =$ Menge der $\zeta(O)$ mit $O \in g$. Man zeigt leicht, daß R^* durch ζ stetig auf $Z(R)$ abgebildet wird.

Hieraus folgt: Unter den Voraussetzung von Satz II ist $Z(R)$ bikompakt.

5.8. Sei $P \in \mathfrak{R}$. Dann ist $\mathfrak{R}^*(P^*) = \mathfrak{R}(P)$.

BEWEIS: Wegen 3.5 ist $\mathfrak{R}(P) \subset \mathfrak{R}^*(P^*)$ evident. Wir brauchen also nur zu einem $g \in \mathfrak{R}^*(P^*)$ ein $a \in \mathfrak{R}(P)$ anzugeben mit $g(a) = g$. $g \prec P \cup (R \setminus \bar{P})$, also gilt für alle $G \in g$:

$$G \not\subset P \cup (R \setminus \bar{P})$$

oder

$$(1) \quad G \cap \mathfrak{R}(P) \neq \emptyset.$$

Wir zeigen, daß die Menge aller

$$(2) \quad \bar{G} \cap \mathfrak{R}(P) \quad \text{mit} \quad G \in g$$

endlich gebunden ist: nach 2.4 hat man zu $G_\kappa \in g$ ($\kappa = 1, \dots, k$) ein $G \in g$ mit

$$G \subset \bigcap_{\kappa=1}^k G_\kappa,$$

also ist

$$\bigcap_{\kappa=1}^k (\bar{G}_\kappa \cap \mathfrak{R}(P)) = \overline{\left(\bigcap_{\kappa=1}^k G_\kappa \right) \cap \mathfrak{R}(P)} \supset \bar{G} \cap \mathfrak{R}(P) \neq \emptyset$$

wegen (1). Die Menge (2) ist also endlich gebunden, also wegen der Bikompaktheit von $\mathfrak{R}(P)$ gebunden. Sei

$$a \in \bigcap_{G \in g} (\bar{G} \cap \mathfrak{R}(P)).$$

Dann ist erstens $a \in \mathfrak{R}(P)$ und zweitens $a \in \bar{G}$ für alle $G \in g$, also auch $a \in G$ für alle $G \in g$, also $g = g(a)$, w.z.b.w.

5.9. Sei R semibikompakt. Dann ist $R^* \setminus R$ nulldimensional.

BEWEIS: Sei $g \in R^* \setminus R$ und O^* eine Umgebung von g . Es gibt ein $P \in \mathfrak{R}$, $P \in g$, $P^* \subset O^*$. Anwendung von 5.8 zeigt: In $R^* \setminus R$ besitzt die Umgebung $P^* \cap (R^* \setminus R)$ von g keinen Randpunkt, ist also ein in O^* enthaltener Brocken rel. $R^* \setminus R$. Also ist $R^* \setminus R$ nulldimensional.

5.10. SATZ IV: Sei R semibikompakt. Dann ist R^* , die Abschließung von R durch seine Endpunkte, bikompakt, ebenso $Z(R)$ der Komponentenraum von R . R ist überall dicht in R^* . Die Menge der Endpunkte $R^* \setminus R$ ist nulldimensional. $Z(R)$ ist nulldimensional.

BEWEIS: Die Bikompaktheit von R^* folgt aus Satz II (wegen 5.4–5 sind die Voraussetzungen erfüllt). Die von $Z(R)$ folgt aus 5.7. Die Dimension von $R^* \setminus R$ folgt aus 5.9; die von $Z(R)$ ergibt sich trivialerweise.

6. Die Kompaktheit von R^*

Wir setzen von R nun immer das zweite Abzählbarkeitsaxiom und Semikompaktheit voraus.

6.1. SATZ V: $Z(R)$ ist dann und nur dann ein Kompaktum, wenn in R jede absteigende Folge nichtleerer Brocken einen nichtleeren Durchschnitt hat.

BEWEIS: Dann: Wegen der Separabilität existiert eine abzählbare Basis

$\mathfrak{U} \subset \mathfrak{B}$ von Z (der Menge der Brocken). Aus dem Mengen von \mathfrak{U} bilde man die Vereinigungen zu je endlich vielen; so entsteht das System \mathfrak{Q} . Sei O ein Brocken. $O = \bigcup_{i=1}^{\infty} U_i$, mit gewissen $U_i \in \mathfrak{U}$. Die Mengen $O \setminus \bigcup_{i=1}^n U_i$ bilden eine absteigende Folgen von Brocken mit leerem Durchschnitt. Nach Voraussetzung ist eine von ihnen leer, also O die Vereinigung *endlich* vieler U_i , also $O \in \mathfrak{Q}$. Also besitzt \mathfrak{D}_β eine *abzählbare* Basis, die besteht aus allen Relationen $Q \subseteq Q$ mit $Q \in \mathfrak{Q}$. Also ist R \mathfrak{D}_β -separabel, also ist $Z(R)$ nach 3.4.3 separabel und nach Satz IV demnach ein Kompaktum.

Nur dann: Gäbe es eine absteigende Folge $M_1 \supset M_2 \supset \dots$ nichtleerer Brocken mit leerem Durchschnitt in R , so wären $N_\nu = M_\nu \setminus M_{\nu+1}$ zueinander fremde Brocken, und ihre Vereinigung wäre der Brocken M . Sei ϕ eine Menge von natürlichen Zahlen und ψ ihre Komplementärmenge. Dann sind die Mengen

$$\bigcup_{\nu \in \phi} N_\nu \quad \text{und} \quad \bigcup_{\nu \in \psi} N_\nu.$$

Komplemente voneinander und als Vereinigungen von offenen Mengen offen, also sind beide Brocken. Also ist

$$\left(\bigcup_{\nu \in \phi} N_\nu \right)^*$$

für jedes ϕ ein Brocken in $Z(R)$; es gibt also in $Z(R)$ Kontinuum viel Brocken, und das widerspricht nach 1.9 der Separabilität von $Z(R)$.

6.2. Sei A abgeschlossen in R und $\mathfrak{R}(A)$ kompakt. Sei $Z(R)$ ein Kompaktum. Dann ist auch $Z(A)$ ein Kompaktum.

BEWEIS: Wir wenden 6.1 an: Sei $M_1 \supset M_2 \supset \dots$ eine Folge von nichtleeren Brocken rel. A . Wir zeigen, daß ihr Durchschnitt nichtleer ist: Sind alle $M_\nu \cap \mathfrak{R}(A) \neq \circ$, so auch ihr Durchschnitt wegen der Kompaktheit von $\mathfrak{R}(A)$, und dann sind wir fertig. Sei also $M_k \cap \mathfrak{R}(A) = \circ$. Sei nun $n \geq k$. M_n ist offen in A , also nun auch offen in $A \setminus \mathfrak{R}(A)$, also auch offen in R . Andererseits war M_n abgeschlossen in A und A abgeschlossen in R , also auch M_n abgeschlossen in R . Die M_n ($n \geq k$) sind also auch Brocken in R , und nach Voraussetzung und 6.1 ist ihr Durchschnitt nichtleer, w.z.b.w.

6.3. SATZ VI: R sei separabel und semikompakt. Dann und nur dann ist die Abschließung R^* von R durch seine Endpunkte ein Kompaktum, wenn der Komponentenraum $Z(R)$ ein Kompaktum ist.

Nur dann folgt aus 5.7, da $Z(R)$ als stetiges Bild von R^* notwendig auch ein Kompaktum sein muß.

Um dann zu beweisen, werden wir für \mathfrak{D}_R eine abzählbare Basis \mathfrak{D}'' konstruieren. Nach Satz I wird R^* dann in der Tat separabel. Unsere Aufgabe ist also die folgende: Wir konstruieren ein abzählbares System $\mathfrak{R}' \subset \mathfrak{R}$ derart, daß zu je zwei Mengen $O, P \in \mathfrak{R}$ mit $O \subseteq P$ eine Menge $Q \in \mathfrak{R}'$ existiert mit $O \subseteq Q \subseteq P$. Gelingt das, so sind wir in der Tat fertig; denn \mathfrak{D}_R war eine Basis von \mathfrak{D}_R , und hat man nun eine Relation $O \subseteq P$ aus \mathfrak{D}_R , so ist definitionsgemäß $O, P \in \mathfrak{R}$; man kann also (die Lösung der "Aufgabe" vorausgesetzt) $O_1, P_1 \in \mathfrak{R}'$

konstruieren mit $O \subset O_1 \subset P_1 \subset P$. Demnach bilden die $O_1 \subset P_1$ mit $O, P \in \mathfrak{R}'$ eine abzählbare Basis \mathfrak{D}'' von \mathfrak{D}_2 .

Die Lösung der Aufgabe geschieht in 6.4–5. Es ist bemerkenswert, daß die Kompaktheit des Komponentenraumes $Z(R)$ eine wesentliche Bedingung für die Lösbarkeit ist.

6.4.1. O_1, O_2, \dots mögen eine Basis von R bilden, $O_n \in \mathfrak{R}$.

6.4.2. \mathfrak{A}_n sei die Menge aller \bar{O}_ν und $R \setminus O_\nu$ mit $\nu \leq n$.

6.4.3. \mathfrak{B}_n sei die Menge aller Durchschnitte, gebildet aus Mengen von \mathfrak{A}_n .

6.4.4. \mathfrak{C}_n sei die Menge aller Brocken aller $B \in \mathfrak{B}_n$. Wegen 6.2 und der Voraussetzung über $Z(R)$ ist auch $Z(B)$ kompakt, also ist die Menge der Brocken jedes B nach 1.9 abzählbar, also ist \mathfrak{C}_n abzählbar.

6.4.5. \mathfrak{R}_n bestehe aus allen Vereinigungen von je endlich viel Mengen aus \mathfrak{C}_n .

6.4.6. \mathfrak{R}'_n bestehe aus den offenen Kernen aller Mengen von \mathfrak{R}_n .

6.4.7. \mathfrak{R}' ist die Vereinigung aller \mathfrak{R}'_n .

Man sieht ohneweiteres, daß $\mathfrak{R}' \subset \mathfrak{R}$ ist.

6.5. Seien $O, P \in \mathfrak{R}$. Gemäß der "Aufgabe" zeigen wir die Existenz eines $Q \in \mathfrak{R}'$ mit $O \subset Q \subset P$.

6.5.1. Wegen der Kompaktheit von $\mathfrak{R}(O)$ gibt es unter den O_1, O_2, \dots der Basis endlich viele, $O_{\nu_1}, \dots, O_{\nu_r}$, die $\mathfrak{R}(O)$ so überdecken, daß $\bar{O}_{\nu_r} \subset P$. Wir setzen $r = \max(\nu_1, \dots, \nu_r)$.

6.5.2. \mathfrak{B}'_r sei die Gesamtheit der $B \in \mathfrak{B}_r$ mit $B \cap O \neq \circ$. \mathfrak{B}'_r ist eine Überdeckung von \bar{O} .

6.5.3. Wir zerlegen \mathfrak{B}'_r in \mathfrak{B}''_r und \mathfrak{B}'''_r , derart daß $B \in \mathfrak{B}''_r$ dann und nur dann, wenn $B \cap \mathfrak{R}(O) \neq \circ$.

6.5.4. Ist $B \in \mathfrak{B}''_r$, so ist einerseits wegen (6.4.2–3) B für jedes $\nu \leq r$ Teilmenge eines $A \in \mathfrak{A}_\nu$, also Teilmenge von \bar{O}_{ν_r} oder $R \setminus O_{\nu_r}$ für jedes ν . Andererseits ist $B \cap \mathfrak{R}(O) \neq \circ$ und wegen 6.5.1 auch $B \cap O_{\nu_r} \neq \circ$ für ein gewisses ν . Zusammen ergibt das: $B \subset \bar{O}_{\nu_r} \subset P$. Also: alle $B \in \mathfrak{B}''_r$ liegen in P .

6.5.5. Sei nun $B \in \mathfrak{B}'''_r$. Dann ist $B \cap \mathfrak{R}(O) = \circ$. Also ist $B \cap O$ eine Brocken in B , also $B \cap O \in \mathfrak{C}_r$. Wir setzen

$$T = \bigcup_{B \in \mathfrak{B}''_r} B \cup \bigcup_{B \in \mathfrak{B}'''_r} (B \cap O).$$

Dann ist $T \in \mathfrak{R}$ (siehe 6.4.5); $T \subset P$ (siehe 6.5.4); $\bar{O} \subset T$ und zwar ist wegen 6.5.1 auch noch jeder Randpunkt von O innerer Punkt von T .

6.5.6. Q sei der offene Kern von T . Dann ist nach 6.4.6: $Q \in \mathfrak{R}'_n \subset \mathfrak{R}'$. Ferner nach 6.5.5: $O \subset Q \subset P$.

Also löst \mathfrak{R}' in der Tat die Aufgabe.

7. Der Charakterisierungssatz

7.1. S sei ein Kompaktum. R genüge den Bedingungen von Satz VI, R sei überall dicht in S , $S \setminus R$ sei nulldimensional. Dann existiert eine stetige Abbildung $f(R^*) = S$, die in R die Identität ist.

BEWEIS: Die in S genommene abgeschlossene Hülle einer Menge werde mit dem Exponenten S angedeutet.

Sei $g \in R^*$. Die Menge aller G^s mit $G \in g$ heie g^s . Da g^s endlich verbunden ist, ist sein Durchschnitt nichtleer; er mge den Punkt $a \in S$ enthalten.

V und W seien Umgebungen von a in S ; $\bar{V} \subset W$; $\mathfrak{R}(V) \subset R$; $\mathfrak{R}(W) \subset R$ (wegen der Nulldimensionalitt von $S \setminus R$ kann man beliebig kleine derartige Umgebungen finden). Fr $G \in g$ ist $a \in G^s$, also $G^s \cap V \neq \circ$, also, da V offen ist, $G \cap V \neq \circ$. Nach 2.3.3 existiert $G_0 \in g$ mit $G_0 \subset W$. Dann ist auch $G_0^s \subset W^s$. Der Durchschnitt aller Mengen von g^s ist demnach in jeder Umgebung W von a enthalten und besteht daher nur aus dem Punkt a , den wir auch $f(g)$ nennen. f bildet nach dem Vorigen jedes $g < G_0$ ab in W^s , also $f(G_0^*) \subset W^s$ (wobei G_0^* von W abhngt und W beliebig klein genommen werden kann). Also ist die Abbildung f stetig. $f(p) = p$ fr $p \in R$ ist evident.

7.2. Die Eigenschaft β' lautet: Ist U eine Umgebung von $a \in S \setminus R$ in S , so ist es unmglich, $U \cap R$ in zwei fremde, in R offene Mengen zu zerlegen, die beide (in S) a als Hufungspunkt besitzen.

7.3. Sei auer den Bedingungen von 7.1 noch β' erfllt. Dann ist die Abbildung f aus 7.1 topologisch.

BEWEIS: Wir brauchen nur die Eineindeutigkeit zu zeigen. Sei $a \in S \setminus R$ und A das f -Urbild von a . A ist abgeschlossen und nulldimensional. Mge A aus mehr als einem Punkt bestehen (wir fhren diese Annahme zum Widerspruch).

Ist die Umgebung (in S) V von a klein genug, so zerfllt $f^{-1}(V)$ in zwei fremde offene Mengen C und D . Ist V' eine Umgebung von a mit $\bar{V}' \subset V$, so zerfllt $f^{-1}(V')$ in zwei fremde offene Mengen, C' und D' , deren abgeschlossene Hllen auch noch fremd zueinander sind, und die, beide, Punkte von A enthalten. Die Zerlegung $C' \cap R$, $D' \cap R$ von $V' \cap R$ widerspricht β' , w.z.b.w.

7.4. $S = R^*$ besitzt die Eigenschaft β' .

BEWEIS: Sei Q eine Umgebung von g in R^* , die gegen β' verstoe und O, P eine gegen β' verstoende Zerlegung von $Q \cap R$. Sei Q' eine Umgebung von g mit $Q' \subset Q$. Dann wird $Q' \cap R$ in $O' = Q' \cap O$ und $P' = Q' \cap P$ so zerlegt, $\bar{O}' \subset O$, $\bar{P}' \subset P$ abgeschlossene Hlle gebildet (in R). g ist Hufungspunkt von O' und P' in R^* . Also gilt fr jedes $G \in g$: $G \cap O' \neq \circ$, $G \cap P' \neq \circ$. Nach 2.3.3 gibt es G_1 und G_2 in g mit $G_1 \subset O$, $G_2 \subset P$. Aus $O \cap P = \circ$ folgt $G_1 \cap G_2 = \circ$ im Widerspruch zu 2.3.1. Es gibt also keinen Versto gegen β' .

7.5. S sei ein Kompaktum. R sei beralldicht in S , und $S \setminus R$ sei nulldimensional und es gelle β' . Dann ist R Semikompaktum und $Z(R)$ Kompaktum.

BEWEIS: Sei $a \in R$. Die Vereinigung von a und $S \setminus R$ ist nulldimensional.⁷ Es gibt also beliebig kleine Umgebungen von a mit zu $S \setminus R$ fremder, also in R kompakter Berandung.

Zum Beweise der Kompaktheit von $Z(R)$ verwenden wir das Kriterium 6.1. Sei $M_1 \supset M_2 \supset \dots$ eine Folge von verschiedenen Brocken aus R . Wir whlen $a_n \in M_n \setminus M_{n+1}$. Wir ziehen aus a_n eine in S konvergente Teilfolge a_{n_n} mit dem Limes a . Wir haben

$$a_{n_n} \in M_{n_n} \setminus M_{n_n+1} \subset M_{n_n} \setminus M_{n_n+1}.$$

⁷ Siehe z. B. K. Menger, *Dimensionstheorie*, Leipzig 1928, S. 115.

Wir setzen $a = a'_n$, $M_{r_n} = M'_n$. Statt $\cap M_n \neq \circ$ können wir auch $\cap M'_n \neq \circ$ beweisen. Wäre das falsch so könnte a nicht in R sein. a wäre dann nach 5.8 innerer Punkt von M_1^* und $U = M_1^*$ wäre Umgebung von a . Dann lieferten

$$V = \bigcup_{n=1}^{\infty} (M_{2n} \setminus M_{2n+1}), \quad W = \bigcup_{n=1}^{\infty} (M_{2n-1} \setminus M_{2n})$$

eine β' widersprechende Zerlegung von $U \cap R$.

7.6. SATZ VII: S sei ein Kompaktum. R sei überalldicht in S . Dann und nur dann ist S wesentlich die Kompaktifizierung von R durch seine Endpunkte, wenn gilt:

$\alpha')$ $S \setminus R$ ist nulldimensional.

$\beta')$ Die Umgebungen eines Punktes $p \in S \setminus R$ werden durch $S \setminus R$ nicht zerlegt in zwei in R offene Mengen, die beide in S den Häufungspunkt p besitzen.

BEWEIS: Dann folgt aus 7.3, wobei 7.5 rechtfertigt, daß wir unter den Voraussetzungen über R die Semikompaktheit von R und die Kompaktheit von $Z(R)$ weggelassen haben. Nur dann folgt aus Satz VI und aus 7.4.

8. Die Endpunkte von Gruppen

8.1. R genüge im Folgenden dem zweiten Abzählbarkeits axiom, sei semikompakt, zusammenhängend und eine topologische Gruppe (die letzte Forderung besagt, daß R eine Gruppe sei, in der $a \cdot b$ eine stetige Funktion von a und b sei und a^{-1} eine stetige Funktion von a); die Identität heiße e .

8.2. Die Links- (oder Rechts-) Multiplikationen mit einem festen Element a sind topologische Abbildungen von R auf sich; nach Satz VII lassen sie sich topologisch bis in die Endpunkte festsetzen. ag ist also sinnvoll als $\lim ac_r$, wenn $\lim c_r = g$; und zwar ist ag wieder ein Endpunkt.

8.3. Seien $O \in \mathfrak{R}$, $P \in \mathfrak{R}$, $\bar{O} \subset P$. Da $\mathfrak{R}(O)$ kompakt und in P ist, so gibt es eine Umgebung U von e mit

$$(1) \quad \mathfrak{R}(aO) = a\mathfrak{R}(O) \subset P \text{ für alle } a \in U.$$

Sei

$$N = \overline{a\bar{O}} \cap (R \setminus P);$$

$$\mathfrak{R}(N) \subset \mathfrak{R}(aO) \cup \mathfrak{R}(R \setminus P),$$

also wegen (1)

$$(2) \quad \mathfrak{R}(N) \subset \mathfrak{R}(P).$$

Seien

$$(3) \quad N_r = \overline{a_r \bar{O}} \cap (R \setminus P),$$

$$(4) \quad \lim a_r = e.$$

Wären alle N_r nichtleer, so wären wegen des zusammenhangs von R auch alle $\mathfrak{R}(N_r)$ nichtleer und es gäbe

$$(5) \quad c_r \in \mathfrak{R}(N_r) \subset \mathfrak{R}(P) \text{ (nach (2)).}$$

Dann nach (3) $c, \in \overline{a, \bar{O}}$, also

$$(6) \quad c, = a, b, \text{ mit } b, \in \bar{O}.$$

Die $c,$ besitzen nach (5) einen Häufungspunkt c in $\Re(P)$, also ist wegen (4) c auch Häufungspunkt der $b,$ und nach (6) in \bar{O} , und das widerspricht der Voraussetzung $\bar{O} \subset P$.

Also ist die Annahme falsch, und es gibt in jeder Folge $a,$ mit $\lim a, = e$ unendlich viel ν mit leerem $N,$, also mit $a, \bar{O} \subset P$. Also gibt es auch eine Umgebung V von e mit

$$a\bar{O} \subset P \text{ für alle } a \in V.$$

Sei der Endpunkt g definiert durch die Folge G_n mit $\overline{G_{n+1}} \subset G_n$.⁸ Dann gibt es nach dem Vorangehenden Umgebungen V_n von e mit $a\overline{G_{n+1}} \subset G_n$ für alle $a \in V_n$. Es ist also $ag < G_n$ für $a \in V_n$.

Sei a_n eine e -Folge. Zu jedem n gibt es ein k_n mit $a, \in V_n$ für $\nu \geq k_n$. Also $a, g < G_n$ für $\nu \geq k_n$. Hier aus folgt

$$\lim a, g = (\lim a,) g$$

(erst für e -Folgen und dann allgemein für konvergente Folgen).

8.4. Die Rechtsmultiplikation mit g bildet also R stetig ab in $R^* \setminus R$. Da R zusammenhängend, $R^* \setminus R$ aber nulldimensional ist, muß die Bildmenge ein Punkt sein, und da $eg = g$ ist dieser Punkt g . Also

$$ag = g.$$

8.5. Daher ist für $\lim c, = g$ und $G \in g$ auch $ac, \in G$ für fast alle ν . Dann gilt aber für jedes kompakte $M \subset R$

$$\text{auch } Mc \subset G \text{ für fast alle } \nu.$$

8.6. Wir fassen die Ergebnisse zusammen in

SATZ VIII: Die Links- (oder Rechts-) Multiplikationen mit Elementen von R sind bis in die Endpunkte hinein topologische Abbildungen. Die Endpunkte sind Fixpunkte dieser Abbildungen. Jede kompakte Menge aus R kann durch Rechts- (und Links-) Multiplikation in jede Umgebung jedes Endpunktes gezogen werden.

Wir beweisen nun

SATZ IX: Eine zusammenhängende, semikompakte, dem zweiten Abzählbarkeitsaxiom genügende Gruppe besitzt höchstens zwei Endpunkte, ist also im Kleinen kompakt.

8.7. BEWEIS: Seien f, g, h drei verschiedene Endpunkte (wir führen diese Annahme zum Widerspruch).

Wir nehmen $F \in f, G \in g, G' \in g, H \in h$, so daß F, G, H paarweise fremd sind und $\bar{G}' \subset G$ ist. Nach Satz V existiert c mit

$$\Re(Fc) = \Re(F)c \subset G'.$$

⁸ Wir nehmen im Folgenden die einen Endpunkt erzeugenden offenen Mengen immer als zu \Re gehörig an.

Setzen wir

$$N = Fc \cap (R \setminus \bar{G}'),$$

so ist (wie in 8.3(2))

$$\mathfrak{R}(N) \subset \mathfrak{R}(G'),$$

also ist N ein Brocken in $R \setminus \bar{G}'$; $f < N$, $h \prec N$ wegen $f < Fc$, $f < R \setminus \bar{G}$, $h \prec Fc$ (Invarianz von f und h bei Multiplikation mit c gemäß Satz VIII). Bildet man $O = N \cap (R \setminus G)$, so sieht man: Es gibt einen Brocken O in $R \setminus G$ mit $f < O$, $h \prec O$. Analog findet man einen Brocken P in $R \setminus G$ mit $h < P$, $f \prec P$. Man darf $O \cap P = \circ$ annehmen (evtl. lasse man aus beiden $O \cap P$ weg).

Der Durchschnitt aller Brocken von $R \setminus G$, die f bzw. h enthalten, heie A_σ bzw. B_σ . Da R zusammenhngend ist, ist $Z(R)$ kompakt, also auch $Z(R \setminus G)$ kompakt (siehe 6.2), also wegen 6.1

$$A_\sigma \neq \circ, \quad B_\sigma \neq \circ,$$

ferner

$$(1) \quad A_\sigma \cap B_\sigma = \circ.$$

$$(2) \quad \text{Fr } G' \subset G \text{ ist } A_{G'} \supset A_\sigma, B_{G'} \supset B_\sigma.$$

Sei $G_1 \supset G_2 \supset \dots$ eine g definierende Folge; wir setzen

$$A = \bigcup_{n=1}^{\infty} A_{G_n}, \quad B = \bigcup_{n=1}^{\infty} B_{G_n}.$$

Wegen (1) und (2) ist

$$(3) \quad A \cap B = \circ.$$

Aus (2) schliet man ohne weiteres, da A und B unabhngig von der Wahl der Folge $G_n \in g$ sind, und hieraus folgt

$$cA_\sigma = A_{c\sigma}, \quad cB_\sigma = B_{c\sigma},$$

also

$$cA = A, \quad cB = B.$$

Das ergibt aber fr

$$c = ba^{-1}$$

einen Widerspruch zu (3), w.z.b.w.

8.8. SATZ X: *Besitzt eine zusammenhngende, semikompakte, dem zweiten Abzhlbarkeitsaxiom gengende Gruppe zwei Endpunkte, so sind die Endpunkte zueinander invers (d.h. ist $\lim c$, ein Endpunkt, so ist $\lim c^{-1}$ der andere. Jede abgeschlossene Untergruppe, die nicht kompakt ist, besitzt dann auch zwei Endpunkte.*

BEWEIS: f und g mögen die Endpunkte sein. g werde durch die Folge $G_1 \supset G_2 \supset \dots$ definiert, $f \prec G_1$. A_σ werde wie oben (8.7) als Durchschnitt der f enthaltenden Brocken definiert. $A = \bigcup_{n=1}^{\infty} A_{\sigma_n}$ ist wieder von der Wahl der G_n unabhängig, also $cA \equiv A$ für alle c , also $A = R$, also $e \in A$.

Wir wählen k so, daß

$$(1) \quad e \in A_{\sigma_k}$$

ist. Dann ist

$$(2) \quad e \in R \setminus G_k.$$

Nach Satz V gibt es zu dem kompakten $\mathfrak{R}(G_1)$ ein $c_n \in G_n$ mit $\mathfrak{R}(c_n G_1) = c_n \mathfrak{R}(G_1) \subset G_k$. Dann ist der Rand von

$$(3) \quad N_n = c_n G_1 \cap (R \setminus G_k)$$

ganz in $\mathfrak{R}(G_k)$, also ist N_n ein Brocken in $R \setminus G_k$; wegen $f \prec G_1$ und der Invarianz der Endpunkte ist $f \prec c_n G_1$, $\prec N_n$, also muß N_n fremd zu A_{σ_k} sein, also wegen (1)

$$e \notin N_n.$$

Hieraus zusammen mit (2) und (3) folgt

$$e \notin c_n G_1$$

oder

$$(4) \quad c_n^{-1} \notin G_1.$$

Der Übergang zur Inversen ist eine topologische Selbstabbildung von R , die sich nach Satz VIII topologisch bis in die Endpunkte fortsetzen läßt. $\lim c_n$ existiert also und ist ein Endpunkt, und nach (4) kann das nicht der Endpunkt g sein. Hieraus folgt dieselbe Aussage für jede Folge c_n mit $\lim c_n = g$, also die erste Hälfte des Satzes.

Ist Q eine abgeschlossene Untergruppe und erzeugen die Folgen G_n und G_n^{-1} die Endpunkte von R , so sind die Folgen $Q \cap G_n$ und $Q \cap G_n^{-1}$ entweder beide leer (und dann ist Q kompakt) oder beide nichtleer (und dann besitzt Q ebenso wie R zwei Endpunkte), und damit ist die zweite Hälfte des Satzes bewiesen.

AMSTERDAM.

ON THE CONTINUATION OF A RIEMANN SURFACE

By MAURICE H. HEINS

(Received November 14, 1941)

1. Introduction.

Let F denote a Riemann surface in the sense of Weyl-Radó [15, 17]. If there exists another Riemann surface G such that F admits a $(1, 1)$ directly conformal map onto a proper part, F' , of G , then F is said to be *continuable* and G is said to be a *continuation* of F ; otherwise, F is said to be *non-continuable* or *maximal*. The closed Riemann surfaces are maximal. Radó [14] has shown by example that there exist open maximal Riemann surfaces. We remark that every open topological surface admits a topological continuation. Later Bochner [1] established by appeal to the well-ordering hypothesis and transfinite induction that every continuable Riemann surface admits a maximal continuation. Shortly thereafter, the question of characterizing the continuable Riemann surfaces was considered by de Possel [10, 11], first, in two notes which are fragmentary in character and do not contain any indication of proofs, and later, in his thesis [12], where he gives a characterization of continuable Riemann surfaces in terms of "sets of maximal type" [12, p. 4] and the topological structure of the surface.

We shall consider the family Φ consisting of Riemann surfaces, G , which are continuations of a given continuable Riemann surface, F , and of F itself. Let it be assumed that F does not admit the Riemann sphere or a closed Riemann surface of genus one as a continuation save when the contrary is mentioned. Under these hypotheses, it will be shown that an explicitly given subset, Φ_0 , of Φ may be defined in a natural manner to be an \mathfrak{L} -space in the sense of Fréchet [3] and that so defined Φ_0 is compact. With the aid of this result, the theorem of Bochner may be established without appeal to transfinite induction. Problems concerning the exhibition of a maximal continuation of a given continuable Riemann surface, and the existence of a maximal continuation with specified properties to be stated in the course of the present paper find an appropriate setting in the study of the structure of Φ and Φ_0 . These questions have not been treated hitherto.

2. Riemann surfaces, Fuchsian and Fuchsoid groups

We recall that a Riemann surface in the sense of Weyl-Radó may be defined as a 2-dimensional manifold F such that to each neighborhood $U(w_0)$ of a point w_0 on the manifold there are associated biuniform and bicontinuous transformations $\tau_{U(w_0)}$ of U onto simply-connected regions of the complex plane with the property that, if U_1 and U_2 are neighborhoods of points w_1 and w_2 of F respectively, and if they have a non-vacuous intersection, then $\tau_{U_2}\tau_{U_1}^{-1}$ defines a $(1, 1)$ directly conformal transformation of $\tau_{U_1}(U_1 \cdot U_2)$ onto $\tau_{U_2}(U_1 \cdot U_2)$ [8, 15, 17].

By virtue of the restriction which we have placed on F —that it admit neither

the Riemann sphere nor a closed Riemann surface of genus one as a continuation — F is not simply-connected. \tilde{F} , the universal covering surface of F , may be defined by its covering relation to F as a Riemann surface. When so defined, \tilde{F} is conformally equivalent to the interior of the unit circle in the complex plane. Thus, being given any point w_0 of F , we may assert the existence of a single-valued map

$$(2.1) \quad w = w(z, w_0)$$

of $|z| < 1$ onto F with the following properties:

1° $w(0, w_0) = w_0$;

2° to each ζ with $|\zeta| < 1$ and $U(\omega)$ where $\omega = w(\zeta, w_0)$ there corresponds a neighborhood of ζ , $N(\zeta)$, such that $\tau_{U(\omega)}[w(z, w_0)]$ is analytic and univalent in $N(\zeta)$;

3° every point w of F has an antecedent with respect to (2.1) in $|z| < 1$;

4° any local determination of the inverse of $w(z, w_0)$, $z(w)$, in $U(\omega)$ for which $\zeta = z(\omega, w_0)$ can be continued along any analytic arc lying in F whose initial point is ω .

Since F is not simply-connected, $w(z, w_0)$ is not univalent and hence by 2°, 3°, and 4° it is automorphic with respect to a group $\mathfrak{G}(F)$ of linear fractional transformations $z \mapsto Tz$ which map $|z| < 1$ onto itself. By 2° the transformations T are either hyperbolic or parabolic, but never elliptic. To avoid circumlocutions, we recall that a group \mathfrak{G} consisting of linear fractional transformations mapping $|z| < 1$ onto itself, which is properly discontinuous for $|z| < 1$, is called *Fuchsian* if it has a finite number of generators, otherwise *Fuchsoid* [2]. Since we shall consider exclusively groups which have no elliptic transformations, we shall employ the unqualified terms “Fuchsian group” and “Fuchsoid group”, understanding that *the groups considered are free from elliptic transformations*.

The group $\mathfrak{G}(F)$ being Fuchsian or Fuchsoid, it follows that to any point z_0 of $|z| < 1$, there corresponds a neighborhood $N(z_0)$ such that every simply-connected region g containing z_0 and contained in $N(z_0)$ satisfies

$$(2.2) \quad T_k g \cdot T_l g = 0 \quad (T_k, T_l \in \mathfrak{G}(F), \quad T_k \neq T_l)$$

We now consider the space F^* formed by identifying the points of $|z| < 1$ which are equivalent with respect to $\mathfrak{G}(F)$ [16, p. 31]. Neighborhoods $U^*(w_0^*)$ of a point w_0^* of F^* are defined to be those sets of points of F^* which correspond to regions of the type g under the identification, where g is to contain a point z_0 of $|z| < 1$ in correspondence with w_0^* . It is readily seen that with this definition F^* is a surface. The maps which associate $U^*(w_0^*)$ with Tg ($T \in \mathfrak{G}(F)$) under the identification define F^* as a Riemann surface. The Riemann surfaces F and F^* are in (1, 1) correspondence and are *conformally equivalent*. We shall denote F^* by $(|z| < 1) \pmod{\mathfrak{G}(F)}$, and, in general, by $E \pmod{\mathfrak{G}}$ we shall denote the space formed by identifying points of a set E of the extended z -plane which are equivalent with respect to a Fuchsian or Fuchsoid group \mathfrak{G} .

Suppose now that we have another map of the type (2.1) of $|z| < 1$ onto F . This second map is automorphic with respect to a Fuchsian or Fuchsoid group $\mathfrak{G}'(F)$. There exists a linear fractional transformation S mapping $|z| < 1$ onto itself such that

$$(2.3) \quad \mathfrak{G}'(F) = S\mathfrak{G}(F)S^{-1}.$$

Conversely, if S is an arbitrary linear fractional transformation mapping $|z| < 1$ onto itself, then the Riemann surface $(|z| < 1) \pmod{S\mathfrak{G}(F)S^{-1}}$ is conformally equivalent to F . Thus *there is associated with F a class of Fuchsian or Fuchsoid groups, each group being obtained from another by taking the transform with respect to an appropriately chosen linear fractional transformation S mapping $|z| < 1$ onto itself, such that when we identify $|z| < 1$ with respect to any one of these groups in the manner indicated above, we obtain a Riemann surface conformally equivalent to F .*

On the other hand, if we start with a class of groups, consisting of a given non-trivial Fuchsian or Fuchsoid group \mathfrak{G} , and its transforms, $S\mathfrak{G}S^{-1}$, where S is an arbitrary linear fractional transformation mapping $|z| < 1$ onto itself, there is associated with this class of groups by the above identification process, a class of conformally equivalent Riemann surfaces, $(|z| < 1) \pmod{S\mathfrak{G}S^{-1}}$. These remarks will be significant in the present discussion.

It may be shown [2, Chap. III] for a Fuchsian or Fuchsoid group \mathfrak{G} , which is not cyclic, that the limit points of the set $\{Tz_0\}$, where z_0 is any point of $|z| < 1$ and $T \in \mathfrak{G}$, consist of either $|z| = 1$ or else of a perfect totally disconnected set of points on $|z| = 1$. In the first case \mathfrak{G} is said to be *of the first kind*, and in the second case *of the second kind*.

3. The families Φ and Φ_0

Let F be a given continuable Riemann surface satisfying the requirement imposed in §1 and let w_0 denote a given point of F . By hypothesis there exists a Riemann surface G and a $(1, 1)$ directly conformal map of F onto a proper subset of G . Let $W = W(w)$ denote this correspondence and let W_0 denote $W(w_0)$. Now let $w(z, w_0)$ and $W(Z, W_0)$ define uniformization maps of $|z| < 1$ and $|Z| < 1$ onto F and G respectively of the type (2.1) such that $w(0, w_0) = w_0$ and $W(0, W_0) = W_0$. The function $Z = f(z)$ which is uniquely defined by the requirements

$$(3.1) \quad f(0) = 0, \quad W[f(z), W_0] = W[w(z, w_0)]$$

has the following properties:

A₁) $Z = f(z)$ is single-valued, analytic and of modulus less than unity for $|z| < 1$.

A₂) If $T \in \mathfrak{G}(F)$, then $f(T) = U_T[f(z)]$ where $U_T \in \mathfrak{G}(G)$.

A₃) If $f(z_1)$ is equivalent to $f(z_2)$ with respect to $\mathfrak{G}(G)$, then z_1 is equivalent to z_2 with respect to $\mathfrak{G}(F)$.

We remark that we may fix $w(z, w_0)$ (which depends upon a single real parameter

θ) once and for all and hence $\mathfrak{G}(F)$. This choice of $w(z, w_0)$ having been made, we may choose $W(Z, W_0)$ so that

$$A_4) f'(0) > 0.$$

Observe that $f(z)$ defines a homomorphism or an isomorphism of $\mathfrak{G}(F)$ into $\mathfrak{G}(G)$. Two cases are to be distinguished: either the image of $\mathfrak{G}(F)$ under this homomorphism is precisely $\mathfrak{G}(G)$, or else the image of $\mathfrak{G}(F)$ is a proper part of $\mathfrak{G}(G)$. This latter situation is illustrated by the example of a Fuchsian group \mathfrak{G} of the second kind. The group \mathfrak{G} is properly discontinuous in the extended z -plane save at a perfect totally disconnected set of points E on $|z| = 1$. If F_1 is the Riemann surface $(|z| < 1) \pmod{\mathfrak{G}}$ and F_2 is the Riemann surface

$$[\text{extended } z\text{-plane deleted in } E] \pmod{\mathfrak{G}}$$

the latter is a continuation of the former. Upon examining the structure of the homomorphism of $\mathfrak{G}(F_1)$ into $\mathfrak{G}(F_2)$ defined by $f(z)$ of (3.1), we find that the image of $\mathfrak{G}(F_1)$ is a proper subgroup of $\mathfrak{G}(F_2)$.

On the other hand, let there be given two groups $\mathfrak{G}:\{T\}$ and $\Gamma:\{S\}$, Fuchsian or Fuchsoid, and let there exist a function $Z = f(z)$ which is analytic and of modulus less than unity, and which satisfies in addition to the condition $f(0) = 0$ the conditions A_1 – A_4 , where \mathfrak{G} replaces $\mathfrak{G}(F)$ and Γ replaces $\mathfrak{G}(G)$. The Riemann surface $G = (|Z| < 1) \pmod{\Gamma}$ is a continuation of the Riemann surface $F = (|z| < 1) \pmod{\mathfrak{G}}$. This follows from A_2 , A_3 , and the analyticity of $f(z)$.

It is to be observed that only those elements S of Γ occur in the homomorphic image of \mathfrak{G} defined by $f(z)$ for which SO is the image of some point of $|z| < 1$ under the map $Z = f(z)$. Hence the strict homomorphic image of \mathfrak{G} , Γ_0 , defined by $f(z)$ also yields a Riemann surface G_0 which is a continuation of F , for $f(z)$, \mathfrak{G} , Γ_0 satisfy the conditions A_1 – A_4 with Γ_0 replacing Γ . We shall denote the class consisting of F and all its continuations by Φ , and by Φ_0 that subclass of Φ consisting of F and all its continuations $\{G_\lambda\}$ for which some $f(z)$ of (3.1) defines a strict homomorphic map of $\mathfrak{G}(F)$ onto $\mathfrak{G}(G_\lambda)$.

It is apparent from the conditions A_1 – A_4 that, if F admits a continuation $G \in \Phi$, it always admits a continuation $G_0 \in \Phi_0$.

4. The family Φ_0 as an \mathcal{L} -space [3]

Let there be given a sequence of elements $\{G_k\}$ ($k = 1, 2, \dots$) and an element G_0 of Φ_0 . The sequence $\{G_k\}$ will be said to have the limit G_0 , denoted by $\mathcal{L}_{k \rightarrow \infty} G_k$, if to each G_k there corresponds a function $f_k(z)$ satisfying the conditions A_1 – A_4 with $\mathfrak{G}(G)$ replaced by $\mathfrak{G}(G_k)$, and if to G_0 there corresponds a function $f_0(z)$ satisfying A_1 – A_4 with $\mathfrak{G}(G)$ replaced by $\mathfrak{G}(G_0)$ such that

$$(4.1) \quad \lim_{k \rightarrow \infty} f_k(z) = f_0(z)$$

for $|z| < 1$, and each f_k and f_0 define strict homomorphic mappings of $\mathfrak{G}(F)$ onto $\mathfrak{G}(G_k)$ and $\mathfrak{G}(G_0)$ respectively. With this definition of a limit, Φ_0 becomes

an \mathfrak{L} -space in the sense of Fréchet. Our first object will be to show that Φ_0 is compact. That is, given any sequence of elements $\{G_k\}$ of Φ_0 , there exists a subsequence $\{G_{n(k)}\}$ and an element G_0 of Φ_0 such that

$$(4.2) \quad \lim_{k \rightarrow \infty} G_{n(k)} = G_0.$$

As a preliminary step in the proof of the compactness of Φ_0 we establish

$$(4.3) \quad \mu = \text{g.l.b. } f'(0) > 0,$$

where $f(z)$ is taken over the class of functions which satisfy the conditions A_1 – A_4 for some $\mathfrak{G}(G)$, $G \in \Phi_0$ (or Φ). If the statement (4.3) were not true, there would exist a sequence of functions $\{f_n(z)\}$ ($n = 1, 2, \dots$) such that $f_n(z)$ satisfies A_1 – A_4 for $\mathfrak{G}(G_n)$ and $\lim_{n \rightarrow \infty} f'_n(0) = 0$. We denote $f'_n(0)$ by α_n , and consider in place of the functions $f_n(z)$, the functions

$$(4.4) \quad \varphi_n(z) = f_n(z)/\alpha_n \quad (n = 1, 2, \dots).$$

From A_1 – A_4 we infer that the function $\varphi_n(z)$ satisfies the following conditions:

B₁) $\varphi_n(z)$ is analytic for $|z| < 1$, $\varphi'_n(0) = 1$.

B₂) If $f_n(T) = U_T^{(n)}[f_n(z)]$, where $T \in \mathfrak{G}(F)$, $U_T^{(n)} \in \mathfrak{G}(G_n)$, then $\varphi_n(T) = V_T^{(n)}[\varphi_n(z)]$, where $V_T^{(n)} = \sigma_n^{-1} U_T^{(n)} \sigma_n$, σ_n being the linear transformation $z \mapsto \alpha_n z$.

B₃) If $\varphi_n(z_1)$ is equivalent to $\varphi_n(z_2)$ with respect to $\sigma_n^{-1} \mathfrak{G}(G_n) \sigma_n$, then z_1 is equivalent to z_2 with respect to $\mathfrak{G}(F)$.

The proper discontinuity of $\mathfrak{G}(F)$ coupled with condition B₃ implies that to each point z_0 of $|z| < 1$ there corresponds a neighborhood $N(z_0)$ such that every member of the sequence $\{\varphi_n(z)\}$ is univalent in $N(z_0)$. In particular, since $\varphi_n(0) = 0$ and $\varphi'_n(0) = 1$, there exists a subsequence, $\{\varphi_{n(m)}(z)\}$, of $\{\varphi_n(z)\}$ which converges uniformly in $N(0)$ and has as its limit function a function univalent in $N(0)$. However, since $\varphi_n(0) = 0$ ($n = 1, 2, \dots$) and $\varphi_n(z)$ is univalent in $N(z_0)$ for each z_0 in $|z| < 1$, the family $\{\varphi_n(z)\}$ is normal for $|z| < 1$ [7, p. 34] and hence by Stieltjes' theorem [7, p. 28], the sequence $\{\varphi_{n(m)}(z)\}$ converges continuously in the sense of Carathéodory for $|z| < 1$. Let $\varphi_0(z)$ denote $\lim_{m \rightarrow \infty} \varphi_{n(m)}(z)$. We note $\varphi_0(0) = 0$, $\varphi'_0(0) = 1$. The linear fractional transformation $V_T^{n(m)}$ converges to the linear fractional transformation V_T , which, *a priori*, is either the identity or else a translation of the Z -plane preserving $Z = \infty$. The condition B₂ yields

$$(4.5) \quad \varphi_0(T) = V_T[\varphi_0(z)];$$

and the condition B₃ and the fact that $\varphi_0(0) = 0$, $\varphi'_0(0) = 1$ imply that, if $\varphi_0(z_1)$ is equivalent to $\varphi_0(z_2)$ with respect to the group $\{V_T\}$, then z_1 is equivalent to z_2 with respect to $\mathfrak{G}(F)$. To see this, we note that, as a consequence of $\varphi_0(0) = 0$ and $\varphi'_0(0) = 1$, $\varphi_0(z)$ is univalent in every $N(z_0)$ where each member of the sequence $\{\varphi_n(z)\}$ is univalent. In particular, $\varphi_0(z)$ maps $N(z_2)$ (1, 1) and directly conformally onto a region in the Z -plane containing $\varphi_0(z_2)$ in its interior. Since

$\varphi_0(z_1)$ and $\varphi_0(z_2)$ are equivalent with respect to $\{V_T\}$, let V denote that member of $\{V_T\}$ for which

$$(4.6) \quad \varphi_0(z_2) = V[\varphi_0(z_1)].$$

Now $V = \lim_{m \rightarrow \infty} V^{(m)}$, where $V^{(m)} \in \mathcal{G}(G_{n(m)})$. Hence for m sufficiently large, $V^{(m)}[\varphi_0(z_1)]$ lies in the image of $N(z_2)$ under $\varphi_0(z)$, $\varphi_0[N(z_2)]$. Since $\varphi_0(z_1) = \lim_{m \rightarrow \infty} \varphi_{n(m)}(z_1)$, $V^{(m)}[\varphi_{n(m)}(z_1)]$ also lies in $\varphi_0[N(z_2)]$ for m sufficiently large. Further, for m sufficiently large and for $\rho(> 0)$ sufficiently small, the circle $N_\rho(\varphi_0(z_2))$ with center $\varphi_0(z_2)$ and radius ρ is covered by the images of $N(z_2)$ with respect to $\varphi_{n(m)}(z)$. The sequence $\{V^{(m)}[\varphi_{n(m)}(z_1)]\}$ has for its limit as $m \rightarrow \infty$, $\varphi_0(z_2)$, and hence for m sufficiently large $V^{(m)}[\varphi_{n(m)}(z_1)]$ lies in $N_\rho(\varphi_0(z_2))$ and hence $V^{(m)}[\varphi_{n(m)}(z_1)]$ has an antecedent $\zeta_{n(m)}$ in $N(z_2)$ with respect to the map $\varphi_{n(m)}(z)$. By B_3 $\zeta_{n(m)}$ is equivalent to z_1 with respect to $\mathcal{G}(F)$. Replacing $N(z_2)$ by a sequence of neighborhoods of z_2 , $\{M_p(z_2)\}$, $M_p(z_2) \subset N(z_2)$, ($p = 1, 2, \dots$) where the diameter of $M_p \rightarrow 0$ as $p \rightarrow \infty$, we find by repeating the above argument

$$(4.7) \quad \lim_{m \rightarrow \infty} \zeta_{n(m)} = z_2.$$

But $\zeta_{n(m)}$ is equivalent to z_1 with respect to $\mathcal{G}(F)$. The proper discontinuity of $\mathcal{G}(F)$ implies $\zeta_{n(m)} = z_2$ for m sufficiently large. In other words, z_2 is equivalent to z_1 with respect to $\mathcal{G}(F)$ as we wished to prove.

The local univalence of $\varphi_0(z)$ guarantees the proper discontinuity of the group $\{V_T\}$ in the finite Z -plane. Suppose, for example, that $\{V_T\}$ contained a sequence $\{V_k\}$ ($V_k \neq Z$, $k = 1, 2, \dots$) which converged to Z itself. Consider any neighborhood of $z = 0$, $N(0)$; its image under $\varphi_0(z)$ would contain a set of points $\{Z_k\}$ ($Z_k \neq 0$) equivalent to $Z = 0$ with respect to $\{V_T\}$ which has $Z = 0$ as a limit point. The antecedents of Z_k in $N(0)$ would then have $z = 0$ as a limit point and would be equivalent to $z = 0$ with respect to the group $\mathcal{G}(F)$. This contradicts the proper discontinuity of $\mathcal{G}(F)$.

The group $\{V_T\}$ must necessarily be one of the following: the identity, a simply-periodic group of translations, a doubly-periodic group of translations.

The system of relations (4.5) and the modified statement of B_3 with φ_0 replacing φ_n and $\{V_T\}$ replacing $\mathcal{G}(G_n)$, together with the stated restrictions on $\{V_T\}$, imply that F admits the Riemann sphere or a closed Riemann surface of genus one as a continuation. But this is contrary to our assumption on F (cf. §1). Hence (4.3) is valid.

The compactness of Φ_0 now follows. For let us consider a sequence of functions $\{f_n(z)\}$ where $f_n(z)$ ($n = 1, 2, \dots$) correlates the Riemann surfaces F and G_n in accordance with A_1 – A_4 , $f_n(z)$ replacing $f(z)$ and $\mathcal{G}(G_n)$ replacing $\mathcal{G}(G)$. The sequence $\{f_n(z)\}$ is bounded by unity and is hence normal for $|z| < 1$. To each point z_0 of $|z| < 1$ there corresponds a neighborhood, $N(z_0)$, in which $f_n(z)$ is univalent ($n = 1, 2, \dots$). There exists a subsequence $\{f_{n(m)}(z)\}$ which converges continuously for $|z| < 1$ as $m \rightarrow \infty$. Denote $\lim_{m \rightarrow \infty} f_{n(m)}(z)$ by

$f_0(z)$. The requirements A_1 and A_4 are obviously satisfied by $f_0(z)$. Since $f_{n(m)} \rightarrow f_0$, $U_T^{n(m)} \rightarrow U_T^0$, where, *a priori*, U_T^0 is either a linear fractional transformation mapping $|Z| < 1$ onto itself, or else a constant of modulus unity. This latter possibility is to be excluded since $f_0(0) = 0$. The set $\{U_T^0\}$ constitutes a group, \mathfrak{G}_0 , since it is the homomorphic image of $\mathfrak{G}(F)$. Finally, by virtue of (4.3) $f_0(z)$ is univalent in the $N(z_0)$ specified above for each z_0 of $|z| < 1$. Hence the group \mathfrak{G}_0 must be properly discontinuous for $|Z| < 1$, if the proper discontinuity of $\mathfrak{G}(F)$ is not to be violated. Obviously \mathfrak{G}_0 contains no elliptic transformations. The requirements A_2 and A_3 are met by $f_0(z)$ and \mathfrak{G}_0 , the proof that A_3 is satisfied being analogous to the proof that $\varphi_0(z)$ satisfies B_3 with $\{V_T\}$ replacing $\sigma_n^{-1}\mathfrak{G}(G_n)\sigma_n$. But then the Riemann surface $G_0 \equiv (|Z| < 1) \pmod{\mathfrak{G}_0}$ is a continuation of F . We have

THEOREM 4.1: *The \mathfrak{L} -space Φ_0 is compact.*

The theorem of Bochner [1, p. 415] is an easy corollary of Theorem 4.1. By the compactness of Φ_0 , there exists a function $f_0(z)$ and a Fuchsian or Fuchsoid group \mathfrak{G}_0 such that A_1 - A_4 are satisfied with $f \equiv f_0$ and $\mathfrak{G}(G) \equiv \mathfrak{G}_0$, and $f'_0(0) = \mu$. If the Riemann surface $(|Z| < 1) \pmod{\mathfrak{G}_0}$, which is a continuation of F , were not maximal, there would exist a function $h(z)$ and a Fuchsian or Fuchsoid group \mathfrak{H} fulfilling the requirements A_1 - A_4 with $f \equiv h$, $\mathfrak{G}(G) \equiv \mathfrak{H}$ and $h'(0) < 1$. Obviously we may take \mathfrak{H} as the homomorphic image of \mathfrak{G}_0 under the mapping defined by $h(z)$. Consider the function

$$(4.8) \quad g(z) \equiv h[f_0(z)]$$

which defines a homomorphic mapping of $\mathfrak{G}(F)$ onto \mathfrak{H} and satisfies in conjunction with \mathfrak{H} A_1 - A_4 , where the appropriate replacements are made. Hence $(|Z| < 1) \pmod{\mathfrak{H}}$ belongs to Φ_0 and is a continuation of F as well as of $(|Z| < 1) \pmod{\mathfrak{G}_0}$. But $g'(0) = h'(0)f'_0(0) < \mu$. This is contrary to the definition of μ . The theorem of Bochner follows.

THEOREM 4.2: *Every continuable Riemann surface admits a maximal continuation.*

We note in passing that the F satisfying the condition of §1 always admit maximal continuations belonging to Φ_0 .

5. Necessary and sufficient conditions that F be continuable

Let F denote a continuable Riemann surface satisfying the restrictions laid down in §1. Up to the present we have made use of the functions $f(z)$ associated with the continuations of F to establish the compactness of Φ_0 . However, we have considered only the grossest properties of $f(z)$ without studying the relation of the structure of the homomorphism defined by $f(z)$ of $\mathfrak{G}(F)$ into $\mathfrak{G}(G)$ to the nature of the continuations which F admits. We now propose to investigate this structure in greater detail.

It is, *a priori*, evident that either for every continuation G of F and every associated $f(z)$ the mapping of $\mathfrak{G}(F)$ into $\mathfrak{G}(G)$ is strictly isomorphic, or else there exists a continuation G^* of F and an associated $f(z)$ such that some element

of $\mathfrak{G}(F)$ other than the identity is mapped into the identity of $\mathfrak{G}(G^*)$. If the first case occurs, then every $f(z)$ defines a (1, 1) map of $|z| < 1$ onto a subregion of $|Z| < 1$. This is a consequence of A_3 which states that

$$f(z_1) = f(z_2)$$

implies z_1 is equivalent to z_2 with respect to some element T , not the identity, of $\mathfrak{G}(F)$. But this would imply

$$f(Tz) = f(z),$$

which is contrary to our assumption on F .

If the second possibility is realized, then a normal subgroup, g , not reducing to the identity, of $\mathfrak{G}(F)$ is mapped on the identity of $\mathfrak{G}(G^*)$ and $f(z)$ is automorphic with respect to g . Let R denote the image in $|Z| < 1$ of $|z| < 1$ under the map $Z = f(z)$. Any determination of the inverse of $f(z)$, say the element at $Z = 0$ which carries $Z = 0$ into $z = 0$, can be continued analytically throughout R by virtue of A_3 . The inverse of $f(z)$ takes on each value of $|z| < 1$ precisely once. If R were simply-connected, the inverse of $f(z)$ would be single-valued in R , and this would contradict the fact that $f(z)$ is automorphic with respect to a group g not the identity. Hence R is a multiply-connected region lying in $|Z| < 1$ and $Z = f(z)$ is a uniformization mapping which defines $|z| < 1$ as the universal covering surface of R .

It may be possible to express the boundary of R , to be denoted by βR , as the sum of two disjoint closed sets σ_1 and σ_2 , one of which, say σ_1 , lies wholly in $|Z| < 1$ and is totally disconnected. If this is so, there exists a closed Jordan curve c lying wholly in R with diameter as small as we please, such that the interior of c contains points of βR , and no two distinct points of the closure of the interior of c are equivalent with respect to $\mathfrak{G}(G^*)$. If this is not the case, then there exists as a consequence of the multiple connectivity of R a continuum K lying wholly in $|Z| < 1$, which is a maximal connected component of βR . This continuum K does not contain a pair of points Z_1 and Z_2 with the property that $Z_2 = UZ_1$, where $U (\neq I) \in \mathfrak{G}(G^*)$. Otherwise, K being maximal,

$$(5.1) \quad UK = K,$$

and, U being either hyperbolic or parabolic, K would not be at a positive distance from $|Z| = 1$. Thus there exists a positive number ϵ for which the ϵ -neighborhood of K contains no two distinct points which are equivalent with respect to $\mathfrak{G}(G^*)$, and consequently there exists a closed Jordan curve k lying in the intersection of R and the ϵ -neighborhood of K , such that K is in the interior of k .

When it is noted that the region R is invariant under the transformations of $\mathfrak{G}(G^*)$, the significance of the existence of the closed Jordan curves c or k becomes clear. For it is readily verified that F is conformally equivalent to $R \pmod{\mathfrak{G}(G^*)}$. The choice of c or k was such that no two distinct points in the

closure of the interior of c or k were equivalent with respect to $\mathfrak{G}(G^*)$. Hence the set of points in $R \pmod{\mathfrak{G}(G^*)}$

$$(5.2) \quad \sum_{U \in \mathfrak{G}(G^*)} R \cdot [\text{interior } c \text{ or } k] \pmod{\mathfrak{G}(G^*)}$$

is conformally equivalent to the intersection of R and the interior of c or k , and is bounded in part by

$$(5.3) \quad \gamma = \sum_{U \in \mathfrak{G}(G^*)} U(c \text{ or } k) \pmod{\mathfrak{G}(G^*)},$$

which is a closed Jordan curve lying in $R \pmod{\mathfrak{G}(G^*)}$. The retrosection γ divides $R \pmod{\mathfrak{G}(G^*)}$ into two disjoint open connected sets one of which is (5.2). This implies that $R \pmod{\mathfrak{G}(G^*)}$ and hence F which is conformally equivalent have boundary elements of the first type [5, p. 164]. We have

THEOREM 5.1: *A necessary condition that there exists a continuation G^* of F such that some $f(z)$ of (3.1) defines a homomorphism of $\mathfrak{G}(F)$ into $\mathfrak{G}(G^*)$ carrying a normal subgroup g (\supset but $\neq I$) into the identity of $\mathfrak{G}(G^*)$ is that F have a boundary element of the first type.*

The converse of this theorem, which is also true, may be demonstrated by a device due to de Possel [12, p. 17]. If F has a boundary element of the first type, then there exists a retrosection γ of F such that $F - \gamma$ is the sum of two disjoint open connected subsets of F , F_1 and F_2 , one of which, say F_1 , is planar (i.e. homeomorphic to a plane region) and multiply-connected. As a consequence of a theorem due to Koebe [6], F_1 may be mapped (1, 1) and directly conformally onto a plane multiply-connected region \mathfrak{F}_1 lying in the interior of the unit circle $|t| = 1$ in the t -plane, such that under this map γ and $|t| = 1$ are in (1, 1) continuous correspondence. The Riemann surface F^* formed from F and $|t| < 1$ by identifying the corresponding points of F_1 and \mathfrak{F}_1 is a continuation of F . To show that some associated $f(z)$ of (3.1) defines a homomorphism of $\mathfrak{G}(F)$ into $\mathfrak{G}(F^*)$ carrying a normal subgroup g (\supset but $\neq I$) into the identity of $\mathfrak{G}(F^*)$, observe that there exists a closed Jordan curve Γ lying in \mathfrak{F}_1 , which contains in its interior a part of the boundary of \mathfrak{F}_1 . The image of Γ in F cannot be deformed to a point in F , whereas the image of Γ in F^* can be deformed to a point in F^* . Hence our assertion follows.

THEOREM 5.2: *If F has a boundary element of the first type, then F admits a continuation F^* such that some associated homomorphic mapping of $\mathfrak{G}(F)$ into $\mathfrak{G}(F^*)$ carries a normal subgroup of $\mathfrak{G}(F)$, g (\supset but $\neq I$), into the identity of $\mathfrak{G}(F^*)$.*

Let us now consider necessary and sufficient conditions that a Riemann surface F admit a continuation G such that some $f(z)$ of (3.1) defines an isomorphism of $\mathfrak{G}(F)$ into $\mathfrak{G}(G)$. We shall give two such conditions, one in terms of the structure of $\mathfrak{G}(F)$, the other in terms of intrinsic properties of F .

Recall that, if an $f(z)$ of (3.1) defines an isomorphic mapping of $\mathfrak{G}(F)$ into $\mathfrak{G}(G)$, then $f(z)$ defines a (1, 1) map of $|z| < 1$ onto a subregion of $|Z| < 1$. Under these circumstances, $\mathfrak{G}(F)$ must be of the second kind. Contrary to

this assumption, $\mathfrak{G}(F)$ would be of the first kind—that is, the set of points $\{\zeta\}$ defined by

$$(5.4) \quad T\zeta = \zeta \quad (T \in \mathfrak{G}(F))$$

would be dense on $|z| = 1$. If T is parabolic, let κ denote an oricycle lying in $|z| < 1$ save for the point ζ and tangent to $|z| = 1$ at ζ ; and if T is hyperbolic, let κ denote a circular arc lying in $|z| < 1$ except for its endpoints ζ and ξ (the other fixed point of T). By A_2 when z tends to ζ along κ , $f(z)$ tends to a fixed point η of U_T , which is the unique fixed point of U_T , if U_T or T is parabolic (in this latter case it may be shown that U_T is then necessarily parabolic); otherwise, $f(z)$ tends to that fixed point of U_T which is $\lim_{|n| \rightarrow \infty} U_T^n$ the $\{n\}$ being the positive or negative integers depending upon whether the n are positive or negative in the relation

$$\lim_{|n| \rightarrow \infty} T^n = \zeta.$$

It follows from a lemma due to Carleman [9, p. 65] that, when z tends radially to ζ , $f(z)$ tends to the fixed point of U_T prescribed above.

Since we are assuming that G is a non-trivial continuation of F (i.e. G is not conformally equivalent to F), the boundary of R , where R is the image of $|z| < 1$ under $f(z)$ must contain points in $|Z| < 1$ and hence points in $|Z| < 1$ which are accessible from R . Let ω_z denote an accessible boundary point of R in $|Z| < 1$, and let α_z denote a Jordan arc connecting $Z = 0$ and ω_z , lying in R save for the endpoint ω_z . The antecedent with respect to $f(z)$ of α_z with ω_z deleted is such that when Z tends to ω_z on α_z , its antecedent tends to a unique point, ω_z , on $|z| = 1$. Application of Carleman's lemma shows that, as z tends to ω_z radially, $f(z)$ tends to ω_z . On the assumption that $\mathfrak{G}(F)$ is of the first kind, there exists a sequence of arcs $\{\beta_n\}$ lying on $|z| = 1$ where

$$\beta_n \supsetneq \beta_{n+1} \quad (n = 1, 2, \dots),$$

each β_n contains ω_z , the endpoints, $\zeta_n^{(1)}$ and $\zeta_n^{(2)}$,

of β_n are fixed points of transformations of $\mathfrak{G}(F)$, and the length of β_n tends to zero monotonically. For each positive integer n , the point ω_z is contained in the Jordan region, σ_n , which is defined by the following two properties:

1⁰ It is bounded by the images with respect to $f(z)$ of the radii joining $z = 0$ to the points $\zeta_n^{(1)}$ and $\zeta_n^{(2)}$ respectively and one of the two arcs of $|Z| = 1$ with endpoints the fixed points of transformations of $\mathfrak{G}(G)$ associated with $\zeta_n^{(1)}$ and $\zeta_n^{(2)}$ in the manner indicated above.

2⁰ σ_n contains the image with respect to $f(z)$ of the region bounded by the radii joining $z = 0$ to $\zeta_n^{(1)}$ and $\zeta_n^{(2)}$ respectively and the arc β_n .

Consider the maximal connected component containing ω_z of the intersection of the boundary of R and the set of points $C: |Z - \omega_z| \leq \rho$, where ρ is positive and is chosen so small that C lies in $|Z| < 1$; denote this component by Ω . I say that $\Omega \subset \sigma_n$ ($n = 1, 2, \dots$). This is a consequence of the fact that $\omega_z \in \sigma_n$ ($n = 1, 2, \dots$). Since the set Ω obviously does not consist solely of the

point ω_z , and since the set of points of the boundary of R which are accessible from R are dense on the boundary of R , the set Ω contains a point $\omega'_z (\neq \omega_z)$ which is accessible from R . Repeating the argument employed before in considering ω_z , we find that there exists a point ω'_z on $|z| = 1$ such that, when z tends to ω'_z radially, $f(z)$ tends to ω'_z . But the point ω'_z must lie on the arc β_n for every positive integer n ; and since the length of β_n tends to zero as $n \rightarrow \infty$, ω'_z must necessarily coincide with ω_z . The contradiction is manifest. Hence $\mathfrak{G}(F)$ is of the second kind.

The converse is also true. Let F be a Riemann surface with $\mathfrak{G}(F)$ of the second kind and let \mathfrak{E} denote the set of points on $|z| = 1$ where $\mathfrak{G}(F)$ is not properly discontinuous. The Riemann surface,

$$(5.5) \quad G = [\text{Extended } z\text{-plane} - \mathfrak{E}] \pmod{\mathfrak{G}(F)},$$

is a continuation of F . It is very simple to exhibit an $f(z)$ of (3.1) associated with this continuation. Observe that the region P consisting of the extended z -plane deleted in the set \mathfrak{E} is a smooth, unbounded, regular covering surface of G . Let $W(z)$ denote the mapping of P onto G defined by the identification (5.5), and let $z = \varphi(Z)$ denote the uniformization mapping of $|Z| < 1$ onto P , where $\varphi(0) = 0$ and $\varphi'(0) > 0$. Then $W[\varphi(Z)]$ defines $|Z| < 1$ as the universal covering surface of G . Observe that $W(z)$ defines $|z| < 1$ as the universal covering surface of $(|z| < 1) \pmod{\mathfrak{G}(F)}$ which is conformally equivalent to F . Now, proceeding as before, we see that $f(z)$ may be defined by

$$(5.6) \quad W(z) = W[\varphi(f(z))]$$

where $f(0) = 0$. This implies, since $\varphi(0) = 0$, that the relation

$$(5.7) \quad z = \varphi[f(z)]$$

holds for $|z| < 1$. The univalence of z itself implies the univalence of $f(z)$. Hence the mapping of $\mathfrak{G}(F)$ into $\mathfrak{G}(G)$ defined by $f(z)$ is isomorphic. In résumé, we have

THEOREM 5.3: *A necessary and sufficient condition that a Riemann surface F admit a continuation G with the property that an associated $f(z)$ of (3.1) defines an isomorphic mapping of $\mathfrak{G}(F)$ into $\mathfrak{G}(G)$ is that $\mathfrak{G}(F)$ be of the second kind.*

We may also characterize Riemann surfaces F for which $\mathfrak{G}(F)$ is of the second kind intrinsically in terms of F . Recall that a transversal σ of an open surface F is a (1,1) continuous image lying in F , $w(t)$, of the open unit interval $(0 < t < 1)$ and having the property that, whenever the sequence $\{t_k\}$ tends to zero or one, the sequence $\{w(t_k)\}$ does not have any limit point in F . We shall agree to call a transversal a τ -transversal if the following conditions are fulfilled:

1^o It separates F in such a manner that one of the connected components, \mathfrak{R} , is simply-connected.

2^o When \mathfrak{R} is mapped (1, 1) and directly conformally onto the interior of the unit circle, $|\zeta| = 1$, the length of the arc of $|\zeta| = 1$ corresponding to τ is less than 2π .

It is readily seen that the existence of a τ -transversal on a Riemann surface F is a conformally invariant property of F .

We now propose to show

THEOREM 5.4: *A necessary and sufficient condition that $\mathfrak{G}(F)$ be of the second kind is that F admit a τ -transversal.*

The necessity follows at once. Let $e^{i\theta}$ be a point of $|z| = 1$ where $\mathfrak{G}(F)$ is properly discontinuous and let κ denote the intersection of $|z| < 1$ with the circumference of a circle with center $e^{i\theta}$ and radius chosen so small that $T_{m\kappa} \cdot T_{n\kappa} = 0$ ($T_m, T_n \in \mathfrak{G}(F)$; $T_m \neq T_n$). Then

$$(5.8) \quad \sum_{T \in \mathfrak{G}(F)} T_{\kappa} \pmod{\mathfrak{G}(F)}$$

is a τ -transversal of $(|z| < 1) \pmod{\mathfrak{G}(F)}$ and this latter Riemann surface is conformally equivalent to F .

The sufficiency of Theorem 5.4 may be demonstrated as follows. If F admits a τ -transversal, the antecedent of \mathfrak{R} in $|z| < 1$ with respect to the uniformization mapping of $|z| < 1$ onto F consists of the sum of disjoint simply-connected regions. We prefer one of these, say r_0 . Since r_0 is in $(1, 1)$ directly conformal correspondence with \mathfrak{R} under the uniformization mapping of $|z| < 1$ onto F by virtue of the monodromy theorem, and since \mathfrak{R} is in $(1, 1)$ directly conformal correspondence with $|\zeta| < 1$ in the manner indicated above, by composing these two mappings, we find that there exists a univalent analytic function, $z(\zeta)$, which maps $|\zeta| < 1$ onto r_0 in such a manner that, whenever ζ tends to any interior point of the arc of $|\zeta| = 1$ complementary to the arc corresponding to τ , $|z(\zeta)|$ tends to unity. Application of Schwarz's reflection principle [4, p. 45] yields the conclusion that $z(\zeta)$ can be continued analytically across the interior of the arc of $|\zeta| = 1$ complementary to the arc corresponding to τ . So continued $z(\zeta)$ maps a sufficiently small neighborhood N of an interior point of the complementary arc $(1, 1)$ and directly conformally onto a region in the z -plane containing a point of $|z| = 1$ in its interior. We take the neighborhood N to be the interior of a circle with its center the point of the complementary arc in consideration. Hence the intersection of N and $|\zeta| < 1$ is mapped by $z(\zeta)$ onto a region on the z -plane lying in r_0 and bounded in part by an arc of $|z| = 1$. But r_0 was so defined that no two distinct points of r_0 are equivalent with respect to $\mathfrak{G}(F)$. Therefore there are points on $|z| = 1$ where $\mathfrak{G}(F)$ is properly discontinuous. The group $\mathfrak{G}(F)$ must be of the second kind.¹

¹ de Possel has introduced in his thesis [12, p. 15] a concept closely related to our τ -transversals. It is that of a simply-connected region D on F which is bounded in part by a finite number or denumerable infinity of transversals of F ; D has the property that, when it is mapped $(1, 1)$ and directly conformally onto the interior of the unit circle, the arcs of the circumference of the unit circle corresponding to the transversals form a set which is not of *maximal type* (i.e.). We have preferred, however, the concept of τ -transversal to that of the regions D because the former concept is apparently the more primitive and simpler.

The following theorem follows readily from the preceding four.

THEOREM 5.5: *If $\mathfrak{G}(F)$ is of the second kind, then F is always continuable. If $\mathfrak{G}(F)$ is of the first kind, then F is continuable if and only if F admits boundary elements of the first type.*

We remark that, if $\mathfrak{G}(F)$ is of the first kind and if F is continuable, then the boundary elements of the first type of F may be characterized conformally by the fact that F admits no τ -transversals.

6. The exhibition of maximal continuations

If a given continuable Riemann surface F has no boundary elements of the first type, it is easy to exhibit a maximal continuation of F . Indeed, the Riemann surface G defined in (5.5) is a maximal continuation of F . To establish this, it suffices to show that (a) $\mathfrak{G}(G)$ is of the first kind, and (b) G has no boundary elements of the first type (Theorem 5.5).

The assertion (a) may be demonstrated as follows. With the region P consisting of the extended z -plane deleted in \mathfrak{E} (§5) we associated its uniformization mapping, $\varphi(Z)$, which is automorphic with respect to a Fuchsoid group, $\mathfrak{G}(P)$, which is of the first kind since \mathfrak{E} is totally disconnected. Now $\mathfrak{G}(G)$ may be taken to be precisely the properly discontinuous group with respect to which $W[\varphi(Z)]$ is automorphic (§5). Hence, $\mathfrak{G}(P)$ being of the first kind, $\mathfrak{G}(G)$ is also and the assertion (a) follows.

Can G admit boundary elements of the first type? If such boundary elements existed, then G would admit a retrosection γ separating G such that $G - \gamma$ is the sum of two disjoint, open, connected sets, one of which, G_1 , is planar and multiply-connected.

The surface G itself may be expressed as the sum of the following disjoint sets: $1^0 F_1 \equiv (|z| < 1) \pmod{\mathfrak{G}(F)}$; $2^0 F_2 \equiv (1 < |z| \leq \infty) \pmod{\mathfrak{G}(F)}$; 3^0 (the set of points on $|z| = 1$ complementary to \mathfrak{E}) $\pmod{\mathfrak{G}(F)}$. The set 3^0 consists, *a priori*, of a system $\{\sigma_k\}$ (finite or denumerable) of transversals and retrosections of G . It cannot however contain any retrosections of G , since, if this were so, F_1 which is conformally equivalent to F would have boundary elements of the first type.

It is clear that the region G_1 must contain points of both F_1 and F_2 . If $G_1 \subset F_1$, then F_1 would have boundary elements of the first type; similarly, if $G_1 \subset F_2$, F_2 , which homeomorphic to F_1 , would have boundary elements of the first type. Since γ is compact on G , it has points in common with only a finite number of the σ_k . The set G_1 deleted in the points of σ_k in G_1 is the sum (finite or denumerable) of planar regions, g_i , each of which must lie wholly in F_1 or wholly in F_2 , since a point of F_1 cannot be connected to a point of F_2 without crossing some σ_k . Furthermore, each g_i must be simply-connected; otherwise either F_1 or F_2 would have boundary elements of the first type.

Observe that at least one g_i is not compact on G . Otherwise, each g_i would be bounded by a closed Jordan curve consisting of points of γ and of a compact subset of the finite number of the σ_k having points in common with γ , this subset

being independent of the index l . It would follow that the intersection of G_1 with $\sum \sigma_k$ is compact on G . Now G_1 itself is not compact on G ; hence there must exist a $(1, 1)$ continuous image of $(0 \leq t < 1)$, $w(t) (\subset G_1)$, such that, whenever $t_p \rightarrow 1$, $\{w(t_p)\}$ is properly divergent. The arc $w(t)$ must lie wholly in F_1 or wholly in F_2 for t sufficiently near unity. Assume that for $t \geq t_0 > 0$, $w(t) \subset F_1$, the treatment of the other possibility being similar. Each g_l , being assumed compact, for $t \geq t_0$, $w(t)$ must have points in common with an infinite number of distinct g_l and hence must cross γ for values of t arbitrarily near one. This is manifestly impossible since γ is compact on G .

Let g_{i_0} denote therefore a g_l which is not compact on G and which lies in F_1 . The simply-connected region g_{i_0} is bounded in part by an arc of γ having end-points on distinct σ_k . The antecedent of g_{i_0} in $|z| < 1$ with respect to the identification mod $\mathfrak{G}(F)$ consists of the denumerable sum of disjoint simply-connected regions, every one of which is bounded by a closed Jordan curve consisting of points of the antecedent of γ and of $|z| = 1$, this Jordan curve containing an arc of $|z| = 1$ with a point where $\mathfrak{G}(F)$ ceases to be properly discontinuous in its interior, since g_{i_0} is not compact on G . This violates the monodromy theorem, since each component of the antecedent of g_{i_0} cannot contain distinct points which are equivalent with respect to $\mathfrak{G}(F)$, and yet each component of the antecedent of g_{i_0} must contain infinitely many distinct points equivalent with respect to $\mathfrak{G}(F)$ in the neighborhood of the points on its boundary where $\mathfrak{G}(F)$ ceases to be properly discontinuous. We infer

THEOREM 6.1: *If F is continuable and has no boundary elements of the first type, then G of (5.5) defines a maximal continuation of F .*

7. A special type of maximal extension

In accordance with the theorem of Koebe already cited [6] it is possible to map a planar Riemann surface F_0 $(1, 1)$ and directly conformally onto a plane region. Furthermore, it is possible to map a plane region $(1, 1)$ and directly conformally onto a region which is dense in the extended complex plane. This is evident, if the region is simply-connected; on the other hand, if the region is multiply-connected, it follows from a theorem of de Possel [13] which states that any multiply-connected plane region may be mapped $(1, 1)$ and directly conformally onto a plane region bounded by a totally disconnected set (possibly vacuous) and a system of segments parallel to the real axis so that this image region is dense in the extended plane.

Does an arbitrary Riemann surface F admit a maximal continuation G such that there is in G a $(1, 1)$ directly conformal image of F which is dense in G ? This question was proposed to the author by Professor Bochner. As we shall see, it is to be answered in the affirmative. Here, too, we shall assume that F does not admit the Riemann sphere or a closed Riemann surface of genus one as a maximal continuation (§1); the treatment of the former case has been indicated and the proof in the latter case may be readily supplied.

We shall say that a Riemann surface G which is a continuation of a given continuable Riemann surface F and which has a dense subset $(1, 1)$ directly conformally equivalent to F , is a *dense continuation* of F . Let Ψ_0 denote the class of Riemann surfaces G which are dense continuations of F .

The class Ψ_0 is not vacuous. Recall that, if F is continuable, then either F has boundary elements of the first type, or else admits a τ -transversal (§5). In the first case, the existence of a $G \in \Psi_0$ is assured by the proof of Theorem 5.2. In the second case, let d denote the simply-connected subregion of F which is bounded in part by τ . The region d may be mapped $(1, 1)$ and directly conformally onto d_x , consisting of the interior of the unit circle in the $x (= x_1 + ix_2)$ -plane deleted in the set $(0 \leq x_1 < 1)$, in such a manner that the transversal τ corresponds to the circumference $|x| = 1$ deleted at the point $x = 1$. Using the device of de Possel once again to identify points of $|x| < 1$ slit in $(0 \leq x_1 < 1)$ and points of d which correspond under the conformal transformation between the two regions being considered, we obtain a dense continuation of F . Hence Ψ_0 is not vacuous.

If there exists a $G \in \Psi_0$ which is maximal, the problem is settled. Otherwise, let ν denote g.l.b. $f'(\theta_0)$ where θ_0 is the class of $f(z)$ of (3.1) associated with the G of Ψ_0 for which the corresponding map of F into G defines G as a dense continuation of F . Clearly $\nu \geq \mu$ (cf. (4.3)). A remark of importance is that $f(z)$ of (3.1) belongs to θ_0 , if and only if the image of $|z| < 1$ under $f(z)$ is dense in $|Z| < 1$. This implies that $\Psi_0 \subset \Phi_0$.

Let $G_1 \in \Psi_0$ and have the property that there is an associated $f_1(z) \in \theta_0$ for which $f'_1(0) < 3\nu/2$. Such a G_1 and $f_1(z)$ exist. Since G_1 is continuable, we apply the argument just employed for F to G_1 . We let Ψ_1 denote the class of dense continuations of G_1 ; we carry over the uniformization of (3.1) taking $|z_1| < 1$ as the universal covering surface of G_1 , letting $z_1 = 0$ correspond to the point $w_0^{(1)}$ of G_1 which is the image of w_0 of F . The class θ_1 bears the same relation to G_1 that θ_0 bears to F , and ν_1 replaces ν . It is to be noted that $\nu_1 \geq 2/3$; else there would exist an $f \in \theta_0$ for which $f'(0) < \nu$. Proceeding as above, we infer the existence of a $G_2 \in \Psi_1$ and an $f_2(z_1) \in \theta_1$ for which $f'_2(0) < 4\nu_1/3$. Advancing step by step, we apply the same argument to G_2 , letting $(\theta_k, \Psi_k, w_0^{(k)}, z_k, \nu_k)$ for $k = 2$ bear the same relation to this system for $k = 1$ that this system for $k = 1$ bears to $(\theta_0, \Psi_0, w_0, z, \nu)$. The connotation of the symbolism is clear. We observe that $\nu_2 \geq 3/4$ and that there exists an $f_3(z_2) \in \Psi_2$ with $f'_3(0) < 5\nu_2/4$. We carry out this procedure inductively obtaining at the n^{th} stage the set $(\theta_n, \Psi_n, w_0^{(n)}, z_n, \nu_n)$ and $f_{n+1}(z_n) \in \Psi_n$, where $\nu_n > (n+1)/(n+2)$ and $f'_{n+1}(0) < (n+3)\nu_n/(n+2)$.

Now consider the sequence of functions, $\{h_n(z)\}$, defined by the inductive relations

$$(7.1) \quad h_1(z) \equiv f_1(z), \quad h_n(z) \equiv f_n(h_{n-1}(z)) \quad (n \geq 2).$$

The function, $h_n(z)$, defines G_n as a dense continuation of F in accordance with the conditions A₁–A₄ of §3. The sequence $\{h_n(z)\}$ converges continuously in the sense of Carathéodory for $|z| < 1$. This is established by observing that

$$(7.2) \quad h'_n(0) > \mu \quad (n = 1, 2, \dots)$$

in accordance with (4.3) and that

$$(7.3) \quad |h_n(z)| \leq |h_{n-1}(z)|$$

for $|z| < 1$ and $n = 2, 3, \dots$ by Schwarz's lemma. As a consequence of the proof of Theorem 4.1, $h(z)$, the limit function of the sequence, $\{h_n(z)\}$, defines a homomorphism or isomorphism of $\mathfrak{G}(F)$ onto a Fuchsoid group \mathfrak{G}^* with the property that $G^* = (|Z| < 1) \pmod{\mathfrak{G}^*}$ is a continuation of F . If we establish that G^* is a dense continuation of F and that G^* is maximal, then the question raised at the beginning of the present section is to be answered in the affirmative.

LEMMA 7.1: *The Riemann surface G^* is a dense continuation of F .*

Define the sequence $\{h_n^{(k)}(z_k)\}$ by the relations

$$(7.1)^{(k)} \quad h_1^{(k)}(z_k) \equiv f_{k+1}(z_k), \quad h_n^{(k)}(z_k) \equiv f_n[h_{n-1}^{(k)}(z_k)] \quad (k = 1, 2, \dots; n = 2, 3, \dots).$$

The argument applied to $\{h_n(z)\}$ shows that the sequence $\{h_n^{(k)}(z_k)\}$ converges continuously for $|z_k| < 1$ as $n \rightarrow \infty$. The limit function, $h^{(k)}(z_k)$, defines an isomorphism or homomorphism of $\mathfrak{G}(G_k)$ onto \mathfrak{G}^* and G^* is a continuation of G_k ($k = 1, 2, \dots$). We also have the further relations

$$(7.4) \quad h^{(k)}[h_k^{(l)}(z_l)] = h^{(l)}(z_l) \quad \text{for } l < k; \quad h^{(k)}(h_k(z)) = h(z) \quad (k \geq 1).$$

Let G_F^* denote the image of F in G^* defined by the map of F into G^* associated with $h(z)$ in accordance with §3; and, in general, let G_k^* denote the image of G_k in G^* defined by the map of G_k into G^* associated with $h^{(k)}(z_k)$ ($k = 1, 2, \dots$). Taking $k = l + 1$ in the first of the relations (7.4) and $k = 1$ in the second, we infer

$$(7.5) \quad G_F^* \subset G_1^* \subset G_2^* \subset \dots$$

We shall show that

$$(7.6) \quad \lim_{n \rightarrow \infty} G_n^* = G^*.$$

Note that

$$(7.7) \quad \lim_{k \rightarrow \infty} h^{(k)}(z) \equiv z.$$

This follows from the fact that

$$\left. \frac{dh^{(k)}(z)}{dz} \right|_{z=0} = \prod_{i>k} f'_i(0),$$

coupled with the relation $\lim_{k \rightarrow \infty} \prod_{i>k} f'_i(0) = 1$ ($\prod_{k=1}^{\infty} f'_i(0) \geq \mu$). If (7.6) were not true, then there would exist a point $w_a^* \in G^*$, not belonging to G_n^* for

any whole number n . But then the functions, $h^{(k)}(z_k)$ ($k = 1, 2, \dots$), would omit for $|z_k| < 1$ the set of points $\{UZ_a\}$ where Z_a is an antecedent of w_a^* in $|Z| < 1$ and $U \in \mathfrak{G}^*$, and hence the relation (7.7) would be violated.

It remains to be shown that $G^* - G_r^*$ is nowhere dense in G^* . Note that $G^* - G_r^*$ is closed, being the complement of G_r^* in G^* , and that by (7.1) it is equal to

$$(7.8) \quad (G_1^* - G_r^*) + (G_2^* - G_1^*) + (G_3^* - G_2^*) + \dots$$

In accordance with the relations (7.4) $G_1^* - G_r^*$ is nowhere dense in G_1^* , and $G_{k+1}^* - G_k^*$ is nowhere dense in G_{k+1}^* for ($k = 1, 2, \dots$). Hence $G_1^* - G_r^*$ and $G_{k+1}^* - G_k^*$ ($k = 1, 2, \dots$) are nowhere dense in G^* . It follows that $G^* - G_r^*$ is of the first category and hence nowhere dense in G^* since it is closed.

The validity of Lemma 7.1 is established. If G^* is maximal, our original question is settled.

Suppose, therefore, that G^* is not maximal. Then there would exist a dense continuation of G^* , say H . Let $f^*(Z)$ have the corresponding significance for G^* and H that $f_{n+1}(z_n)$ has for G_n and G_{n+1} . Since the Riemann surface H is a dense continuation of G_n ($n = 1, 2, \dots$), it follows that

$$(7.9) \quad f^*[h^{(n)}(z_n)] \in \Theta_n \quad (n = 1, 2, \dots).$$

But ν_n of Θ_n satisfies $\nu_n > (n+1)/(n+2)$, and hence by (7.7)

$$\left. \frac{df^*(Z)}{dZ} \right|_{Z=0} = 1;$$

by Schwarz's lemma $f^*(Z) \equiv Z$, which is manifestly contrary to the assumption made. We have

THEOREM 7.1: *A continuable Riemann surface always admits a maximal dense continuation.*

8. A remark

We have left out of consideration those Riemann surfaces which admit the sphere or a closed Riemann surface of genus one as a continuation in order to preserve the unity of presentation. The treatment of these classes of Riemann surfaces follows from our discussion with appropriate modifications.

THE INSTITUTE FOR ADVANCED STUDY

BIBLIOGRAPHY

1. S. BOCHNER, *Fortsetzung Riemannscher Flächen*, Math. Annalen, vol. 98 (1927), pp. 406-421.
2. L. R. FORD, *Automorphic Functions*, New York, 1929.
3. M. FRÉCHET, *Sur quelques points du calcul fonctionnel*. Thesis, Paris, 1906.
4. G. JULIA, *Principes géométriques d'analyse I*, Paris 1930.
5. B. V. KERÉKJÁRTÓ, *Vorlesungen über Topologie I*, Berlin 1923.
6. P. KOEBE, *Über die Uniformisierung beliebiger analytischer Kurven III*, Göttinger Nachrichten, 1908, pp. 337-358.

7. P. MONTEL, *Leçons sur les familles normales des fonctions analytiques*, Paris 1927.
8. R. NEVANLINNA, *Ein Satz über offene Riemannsche Flächen*, *Annales Acad. Sci. Fenn.*, Ser. A T.54. No. 3. (1940).
9. R. NEVANLINNA, *Eindeutige analytische Funktionen*, Berlin 1936.
10. R. DE POSSEL, *Sur le prolongement des surfaces de Riemann*, *C. R. Acad. Sci. de Paris*, vol. 186 (1927) pp. 1092-1095.
11. R. DE POSSEL, *Sur le prolongement des surfaces de Riemann*, *C. R. Acad. Sci. de Paris*, vol. 187 (1928) pp. 98-100.
12. R. DE POSSEL, *Quelques questions de représentation conforme*. Thesis, Paris 1932.
13. R. DE POSSEL, *Zum Parallelschlitztheorem unendlich-vielfach zusammenhängender Gebiete*, *Göttinger Nachrichten*, 1931, pp. 199-202.
14. T. RADÓ, *Ueber eine nicht-fortsetzbare Riemannsche Mannigfaltigkeit*, *Math. Zeitschrift*, vol. 20 (1924) pp. 1-6.
15. T. RADÓ, *Ueber den Begriff der Riemannsche Fläche*, *Acta Szeged*, vol. 2 (1923) pp. 101-121.
16. H. SEIFERT AND W. THRELFALL, *Lehrbuch der Topologie*, Leipzig 1934.
17. H. WEYL, *Die Idee der Riemannschen Fläche*, Leipzig 1923.

LATTICE-ORDERED GROUPS

By GARRETT BIRKHOFF

(Received December 10, 1941)

1. Introduction

We shall be concerned below with lattice-ordered groups, or *l-groups*,¹ in the sense of the following definition.

DEFINITION. *An l-group is*

(I) *a group, on which is defined a binary inclusion relation which is "homogeneous" in the sense that*

(II) $x \geq y$ *implies* $a + x + b \geq a + y + b$ *for all* a, b ,

and relative to which

(III) *the group is a lattice.*

This defines l-groups as abstract algebras; as such, we can (and shall) apply to them such general algebraic concepts as l-subgroup (subalgebra), isomorphism, homomorphism, and so on.

Three important topics will be included as special cases under the single heading of l-groups: the additive and multiplicative groups of ordered fields, which have long been studied by Hahn, Artin, and others, and are now extensively used in valuation theory; the study of abstract number and ideal theory initiated by Dedekind, and recently amplified by Krull, Ward, Lorenzen, Clifford, Dilworth and others; and the semi-ordered function spaces studied very recently by Riesz, Freudenthal, Kantorovitch, the author, Bohnenblust, Stone, Kakutani, and others.

It should be stressed, however, that up to the present time only l-groups which are commutative or simply ordered have been studied; and it came as a considerable surprise to the author that the non-commutative case involved so few new difficulties.

The material below breaks up rather naturally into several parts. First (§§2-8) Postulates (I)-(III) are discussed, and various other equivalent systems of postulates (together with numerous examples) are derived. Then, after brief preliminaries on algebraic formalism, the general structure and decomposition theory of l-groups is treated (§§9-13). After this, a complete classification of *commutative* l-groups whose structure lattice has finite length is given (§§14-20). After this, in §§21-26, special properties of *complete* l-groups are discussed. Fifth, two important generalizations are taken up (§§27-28). The paper then concludes with a list of sixteen unsolved problems (§§30-31), some of which are fundamental.

¹ We are adopting the convenient terminology of M. H. Stone, "A general theory of spectra. II.," Proc. Nat. Acad. Sci. 27 (1941), pp. 83-87.

2. Explanation of (I)

The reader will be assumed to be familiar with the definitions of a group and of a commutative group, and with the algebraic manipulation of the elements of such groups under the additive notation. Thus 0 will denote the group identity, $-a$ the group inverse of a , $a + b$ the result of combining a with b , and na (n any integer) will denote the n^{th} "power" of a in the cyclic subgroup generated by a .

In the commutative case, the rules of manipulation may be summarized in the statement that the group behaves like a vector space over the domain of integers. As it may be shown that every element of an l-group is of infinite order—that $na = 0$ implies $n = 0$ or $a = 0$,—even the cancellation laws hold. In addition,² an equation of the form $nx = ma$ ($n \neq 0$) has at most one solution. If such a solution exists, it may be denoted $(m/n)a$, and regarded as a rational scalar multiple of a . In particular (*op. cit.*, §1) the correspondence $(m/n)a \rightarrow (m/n)$ is, for any fixed a , and *isomorphism* of the set of rational multiples of a and a subgroup (which always contains all integers) of the additive group of all rational numbers. Thus the set of all rational scalar multiples of a may be thought of as a generalized cyclic subgroup.

3. Explanation of (II)

The concept of homogeneity applies to any binary relation on a group.

DEFINITION. A binary relation \geq on a group G is called left-homogeneous if and only if $x \geq y$ implies, for all $a \in G$, that $a + x \geq a + y$ and right-homogeneous if and only if it implies $x + a \geq y + a$. A relation which is both left- and right-homogeneous is called homogeneous.

THEOREM 1. Homogeneity is equivalent to the assertion that (II') every group translation $x \rightarrow a + x + b$ is a lattice-automorphism. (I)³

PROOF. The condition (II') is clearly sufficient. To prove its necessity, recall first that all group translations are one-one. Second, not only does $x \geq y$ imply $a + x + b \geq a + y + b$, but conversely $a + x + b \geq a + y + b$ implies

$$(-a) + a + x + b + (-b) \geq (-a) + a + y + b + (-b)$$

or $x \geq y$. That is, (II) implies that any group translation is an automorphism with respect to the relation \geq , as asserted.

THEOREM 2. Homogeneity is equivalent to the assertion that every transformation of the form⁴ $x \rightarrow a - x + b$ is a dual automorphism. (I)

² For the special properties of such groups, cf. Reinhold Baer, "Abelian groups without elements of finite order," *Duke Jour.* 3 (1937), pp. 68-122.

³ The postulate numbers in parentheses after the statement of a theorem refer to the postulates which are needed to prove the theorem in question. Many of the theorems below have a generality which far transcends the theory of l-groups.

⁴ Especially interesting are the "inversions" $x \rightarrow a - x + a$, which are of period two and have a for fixpoint.

(This means that the correspondence replaces the given homogeneous relation by its converse.)

PROOF. Assuming (II), $x \geq y$ is equivalent to

$$a + (-x) + x + (-y) + b \geq a + (-x) + y + (-y) + b$$

by Theorem 1. But this is $a - y + b \geq a - x + b$, and so the condition is necessary. It is sufficient since it implies that $x \rightarrow 0 - ((-b) + x + (-a)) + 0 = a - x + b$ is the product of two dual automorphisms, hence an automorphism.

DEFINITION. An element a of an l -group G is called positive if $a \geq 0$. The set of all positive elements of G will be denoted G^+ .

THEOREM 3. Homogeneity is equivalent to the assertion that, for some set of "positive" elements invariant under all inner automorphisms $x \rightarrow -a + x + a$, $x \geq y$ if and only if $x - y$ is positive. (I)

PROOF. Assuming (II), clearly $x \geq y$ if and only if $x - y \geq y - y = 0$; moreover $t \geq 0$ implies $-a + t + a \geq -a + 0 + a = 0$ for all a . Conversely, for any set S invariant under all inner automorphisms, the relation $(x - y) \in S$ is homogeneous since $(a + x + b) - (a + y + b) = -(-a) + (x - y) + (-a)$ is, for all $a, b \in G$, the transform of $x - y$ under an inner automorphism.

COROLLARY. If G is commutative, homogeneity is equivalent to the assertion that, for some set of positive elements, $x \geq y$ if and only if $x - y$ is positive.

4. Explanation of (III)

Postulate III asserts that the inclusion relation $x \geq y$ satisfies the usual conditions,

P₁. For all x , $x \geq x$,

P₂. If $x \geq y$ and $y \geq x$, then $x = y$,

P₃. If $x \geq y$ and $y \geq z$, then $x \geq z$,

L'. Any two elements x and y have a l.u.b. $x \cup y$,

L''. Any two elements x and y have a g.l.b. $x \cap y$.

We recall that in any lattice,⁵ the three relations $x \geq y$, $x \cap y = y$, and $x \cup y = x$ are mutually equivalent; indeed, this is even true in any "partially ordered system" satisfying P₁-P₃. It follows that an automorphism with respect to one of the relation or operations \geq , \cup , \cap is necessarily an automorphism with respect to all three. Hence we get as a corollary of Theorem 1,

THEOREM 4. Left-homogeneity is equivalent to either of the dual left-distributive laws⁶

$$(1) \quad a + (x \cup y) = (a + x) \cup (a + y),$$

$$(1') \quad a + (x \cap y) = (a + x) \cap (a + y).$$

right-homogeneity to either right-distributive law (I, P₁-P₃).

⁵ The terminology and notation are identical with that of the author's book "Lattice theory," New York, 1940, although scant use will be made of the theorems proved there.

⁶ Discovered by Dedekind and independently Freudenthal; see footnote 13.

(In case $L'-L''$ do not hold, the existence of the join (meet) on one side of an equation is intended to be equivalent to the existence of that on the other.)

We can prove from the left- and right-distributive laws just stated, and finite induction, also the following more general finite distributive laws

$$\begin{aligned} a + \vee y_j &= \vee (a + y_j) \quad \text{and} \quad \vee x_i + a = \vee (x_i + a), \\ \vee x_i + \vee y_j &= \vee (x_i + y_j), \\ \sum_i (\vee_j x_{i,j}) &= \vee_{j(i)} (\sum_i x_{i,j(i)}), \end{aligned}$$

and their lattice duals.

THEOREM 5. *Homogeneity is equivalent to the "monotonicity law":*

$$(2) \quad x \geq x' \quad \text{and} \quad y \geq y' \quad \text{imply} \quad x + y \geq x' + y'. \quad (I, P_1, P_3)$$

PROOF. Applying homogeneity twice, we get

$$x + y \geq x + y' \geq x' + y',$$

whence (2) follows by P_3 . Conversely, assuming P_1 , we get as special cases of (2), $x + y \geq x + y'$ and $x + y \geq x' + y$, implying homogeneity.

Again, a permutation of the elements of a partially ordered system is a dual automorphism if and only if it interchanges the operations \cup and \cap . Hence, from Theorem 2, we get

THEOREM 6. *Homogeneity is equivalent to the laws*

$$(3) \quad a - (x \cap y) + b = (a - x + b) \cup (a - y + b),$$

$$(3') \quad a - (x \cup y) + b = (a - x + b) \cap (a - y + b). \quad (I, P_1-P_3)$$

We note as a special case

$$(4) \quad x \cap y = -(-x \cup -y) \quad \text{and dually.}$$

From this we see that the lattice postulate L'' is redundant, in the sense that it is implied by I, II, P_1-P_3 , and L' .

5. Stone's postulates

But now P_1-P_3 and L' are equivalent by pure lattice theory to the assertion that our system admits an idempotent, commutative and associative operation $x \cup y$, in which $x \geq y$ means $x \cup y = x$. Hence splitting (1) in two parts (right- and left-translations), we get as a corollary of Theorem 4 and the redundancy of L'' ,

THEOREM 7 (Stone⁷). *An l-group may be defined as a group, with a second*

⁷ Stone assumed the group to be commutative, in which case one of the distributive laws (1') can be omitted (cf. Stone, *op. cit.*).

binary operation \smile which is idempotent, commutative, and associative, and satisfies the distributive laws

$$(1'') \quad \begin{aligned} a + (x \smile y) &= (a + x) \smile (a + y) \\ (x \smile y) + b &= (x + b) \smile (y + b). \end{aligned}$$

It is a curious fact that, in virtue of the duality principle, substitution of \frown for \smile in the above system of postulates should also define an l-group!

Not only can we delete L'' from our list of postulates, but we can even weaken L' .

DEFINITION. By the positive part a^+ of an element a of an l-group, is meant $a \smile 0$; $a^- = a \frown 0$ is dually called the negative part of a .

Using right-homogeneity, we get

$$(5) \quad a \smile b = (b - a)^+ + a = (a - b)^+ + b.$$

Combining (5) with the dualization law (3), we get

$$(6) \quad a \frown b = -(-a + (a - b)^+) = -(a - b)^+ + a.$$

There follows immediately

THEOREM 8. The lattice hypotheses $L'-L''$ can be replaced by the condition that, for all a , $a \smile 0$ should exist. (I, II, P_1 - P_3)

Also, a subgroup of an l-group is an l-subgroup if and only if it contains the positive part of each of its members. Substituting in Theorem 7, we get a further corollary.

THEOREM 9. An l-group may be defined as a group with a unary operation $a \rightarrow a^*$ which satisfies

$$(7) \quad 0^* = 0, \quad (8) \quad c = c^* - (-c)^*,$$

(9) the operation $(a - b)^* + b$ is associative.

A worth-while problem would be to find a less clumsy form of (9). In this connection, $(a^*)^* = a^*$ might be a useful partial substitute. One might also try setting the middle letter of the associative law equal to 0.

6. Examples

In the next two sections, we shall be using Theorem 3 as our main tool.

First, we note that in order to describe an l-group G up to isomorphism, it is sufficient by Theorem 3 to describe the set G^+ of "positive" elements; indeed, this principle is independent of Postulate III. We shall now describe some important examples of l-groups in this way.

EXAMPLE 1. G is the additive group of real numbers; G^+ consists of all those which are non-negative.

EXAMPLE 2. G is the additive group of the integers; G^+ is defined as in Example 1.

EXAMPLE 3. G is the group of all positive rational numbers under multiplication (the integer one is the group identity); G^+ is the set of all positive integers.

EXAMPLE 4. G is the group of all vectors $x = (x', x'')$ with two real components; G^+ contains x if and only if $x' > 0$, or $x' = 0$ and $x'' \geq 0$.

EXAMPLE 5. G is the additive group of all real functions defined on the interval $0 \leq x \leq 1$; G^+ consists of all those which are non-negative (satisfy $f(x) \geq 0$ for all x).

EXAMPLE 6. G is the additive group of functions of bounded variation on $0 \leq x \leq 1$ with $f(0) = 0$; G^+ defined as in Example 5.

EXAMPLE 7. G as in Example 6; G^+ consists of all "increasing" functions (functions for which $x \geq y$ implies $f(x) \geq f(y)$).

We shall now list some examples of non-commutative l-groups. The simplest example consists of the two-parameter non-Abelian Lie group, lexicographically ordered as follows.

EXAMPLE 8. G consists of all couples (x, y) of real numbers, where addition is defined by the formula

$$(x, y) + (x', y') = (x + x', e^{x'}y + y');$$

G^+ consists of all those couples with $x > 0$ or $x = 0, y \geq 0$.

EXAMPLE 9. G has three generators of infinite order, and defining relations $a + b = b + a, a + c = c + b, b + c = c + a$; G^+ contains $ma + m'b + nc$ if and only if $n > 0$, or $n = 0$ while $m \geq 0$ and $m' \geq 0$.

EXAMPLE 10. G consists of the $x > 0$ of any ordered field or skew-field^{7a} (division ring) under multiplication; G^+ consists of all $x \geq 1$.

For purposes of comparison, we shall also list various other examples which satisfy Postulates I-II and part, but not all, of Postulate III.

EXAMPLE 11. G is any group; G^+ consists of the identity 0 alone.

EXAMPLE 12. G is the multiplicative group of all non-zero elements of any algebraic number field; G^+ is the subset of all (algebraic) integers in G .

EXAMPLE 13. G is the group of all elements of any integral domain of characteristic infinity under addition; G^+ is the subset of all sums of squares.

7. Postulates of order reinterpreted

It is trivial that the systems described in Examples 1-13 satisfy Postulates I-II if $a \geq b$ is defined to mean $(a - b) \in G^+$ (cf. Theorem 3). We shall now give simple tests for the validity of parts P_1 - P_3 of Postulate III.

LEMMA 1. *The reflexive law P_1 is equivalent to (9). The group identity is positive. (I, II)*

For $(a - a) \in G^+$ is equivalent to $0 \in G^+$ by group theory. This condition is evidently satisfied in Examples 1-12 above.

LEMMA 2. *The transitive law P_3 , the monotonicity law (2) of Theorem 5, and the condition that*

(10) *Any sum of positive elements is positive, are mutually equivalent. (I, II, P_1)*

PROOF. By Theorem 5, P_3 implies (5) modulo I, II, P_1 . Again (5) implies

^{7a} Cf. K. Reidemeister, "Grundlagen der Geometrie", p. 40.

the closure of G^+ as the special case $a \geq 0$ and $b \geq 0$ imply $a + b \geq 0 + 0 + 0$. Finally, since $(a - b) + (b - c) = (a - c)$, the closure of G^+ implies P_2 as in Theorem 3.

COROLLARY. *The positive elements of any l-group form a semigroup.**

It is also a corollary that P_2 holds in Examples 1-13 above.

LEMMA 3. *The antisymmetric law P_2 is equivalent to asserting that (11) a and $-a$ are both positive only if $a = 0$. (I, II)*

PROOF. If $(x - y) \in G^+$ and $(y - x) \in G^+$ imply $(x - y) = 0$, then P_2 holds. Conversely, if P_2 holds, $z \geq 0$ and $-z \geq 0$ imply $z = 0$.

It may now be checked easily that P_2 holds in Examples 1-11 above, although not in Examples 12-13.

A similar lemma, irrelevant here, is that the symmetric law ($a \geq b$ implies $b \geq a$) is equivalent to asserting that $a \in G^+$ implies $-a \in G^+$. From this and Lemmas 1-2 it follows that G^+ defines an *equivalence* relation if and only if it is a *subgroup* of G . (I, II)

Finally, we can read off from Lemmas 1-2 and Theorem 8, the following not very satisfactory result.

LEMMA 4. *The lattice hypotheses $L'-L''$ are equivalent to the following condition: (12) Given a , there exists a^+ such that u and $(u - a)$ are both positive if and only if $(u - a^+)$ is. (I, II, P_1 , P_2)*

From Theorem 3 and Lemmas 1-4, we conclude as the final theoretical result of this section,

THEOREM 10. *An l-group may be defined as a group G with a subset G^+ of "positive" elements which satisfies conditions (9)-(12).*

We also conclude that Examples 1-10 above are l-groups. In Examples 1, 2, 4, 8, 10 this is true because the ordering is simple: for all a , either a or $-a$ is in G ; and so a^+ is a or 0 accordingly. In Example 3, $(m/n)^+$ is the numerator of m/n when written in lowest terms; in Examples 5-6, f^+ is the "positive part" of f as usually defined (equal, for all x , to the larger of $f(x)$ or 0); in Example 7, f^+ is the "positive variation" of f . In Example 9, $(ma + m'b + nc)^+$ is 0 if $n < 0$, $(ma + m'b + nc)$ if $n > 0$, and $m^+a + m'^+b$ if $n = 0$.

8. Fifth set of postulates

We have characterized l-groups by four sets of postulates. Our definition was in terms of the group operation and a binary relation; Theorem 7 in terms of the group operation and a binary operation; Theorem 9 in terms of the group operation and a unary operation; Theorem 10 in terms of the group operation and a unary relation or set. We shall now give a fifth set of postulates for l-groups which, oddly enough, is in terms of the group operation alone!

Evidently any l-group or other lattice has the following "Moore-Smith" property:

* By a *semigroup*, we mean a system closed under an associative binary operation and having an identity element, in which the laws of cancellation hold ($a + x = a + y$ implies $x = y$ and so does $x + a = y + a$).

(13) Given a, b , there exists c with $c \geq a$ and $c \geq b$.

LEMMA 1 (Clifford⁹). *The Moore-Smith property is equivalent to the assertion that*

(14) *Every element is a difference of positive elements.* (I, II, P_1 , P_3)

PROOF. Assuming (13) with $b = 0$, we get $a = c - (-a + c)$, where $c \geq 0$ and $-a + c = -a + (c - a) + a \geq -a + 0 + a = 0$. Conversely, if $a = a' - a''$ and $b = b' + b''$, where a', a'', b', b'' are positive, then $c = a' + b'$ exceeds both a and b .

Now let A be any group with a relation \geq satisfying II, P_1 , P_3 and (14). We shall show that A is determined to within isomorphism by the semigroup A^+ of its positive elements. The proof is related to the general theory of the extension of semigroups to groups.

THEOREM 11. *In the notation of the calculus of complexes, $a + A^+ = A^+ + a$, for all $a \in A$.*

PROOF. Both sets consist of the elements containing a .

COROLLARY. *Given a and x in A^+ , there exist a unique $y \in A^+$ such that $a + x = y + a$ and $z \in A^+$ such that $x + a = a + z$.*

The existence follows from Lemma 1; the uniqueness from the cancellation postulate defining semigroups.

Now observe that A consists by (14) of the differences $b - c$ of elements of A , equated and combined by the rules

(15) $b - c = b' - c'$ if and only if t, u exist in A^+ such that $b + t = b' + u$ and $c + t = c' + u$,

(16) $(b - c) + (b' - c') = (b + b') - (c' + c'')$, where c'' is the unique solution of $b' + c'' = c + b'$.

The sufficiency of (15) is clear; as regards the necessity, if we choose $s \geq b, b'$, $t = -b + s, u = -b' + s$, then $b + t = s = b' + u$, while if $b - c = b' - c'$, then $b + t - t - c = b' + u - u - c'$, whence $-t - c = -u - c'$ and $c' + u = c + t$.

Clearly equations (15)-(16) describe the group structure of A in terms of that of A^+ . Moreover since

(17) $(b - c) \in A^+$ if and only if $b = c + t$ for some $t \in A^+$,

the lattice structure of A can also be described in terms of the group structure of A^+ . In fact, $b - c \geq b' - c'$ if and only if t, u exist such that $b + t \geq b' + u$ and $c + t \leq c' + u$.

Conversely, suppose S is any semigroup in which, for all $a, a + S = S + a$ (in multiplicative language, such that the left-multiples of any element are all right-multiples, and conversely). Then equations (15)-(16) may be shown to define a group.¹⁰ We shall omit the details; one shows that (15) defines an

⁹ A. H. Clifford, "Partially ordered Abelian groups," *Annals of Math.* 41 (1940), pp. 465-473, esp. p. 467. From the equation $a = 1/4((a+1)^2 - (a-1)^2)$, we see that (14) holds in Example 13.

¹⁰ R. Baer has proved, in conversation, that a group can be constructed whenever, given a and b , x and y can be found such that $a + x = b + y$.

equivalence relation, which is a congruence relation with respect to the addition defined by (16), and relative to which the latter is associative, and gives any element $b - c$ an inverse $c - b$. Furthermore, under (17), the set of positive elements forms a subset of A isomorphic with S .

It follows, by Theorem 10, that we get an l -group provided (11)–(12) hold. But now (11) is clearly equivalent to

(17') *If $a + b = 0$ in S , then $a = b = 0$.*

Finally, if any two elements of S have a l.u.b. with respect to the definition

(17'') *$b \geq c$ if and only if $b = c + t$ for some $t \in S$,*

then for any $a = a' - a''$ of A ($a', a'' \in S$) there exists $a^+ = (a' \cup a'') - a''$, which proves that condition (12) holds. There follows

THEOREM 12 (von Neumann¹¹). *An l -group may be defined as the extension to a group of a (multiplicative) semigroup S , in which (i) $ab = 1$ implies $a = b = 1$, (ii) $aS = Sa$ for all a , (iii) any two elements have a least common multiple. In this group S consists of the positive elements.*

COROLLARY¹². *A commutative l -group may be defined as the extension to a group of a commutative semigroup S , in which (i) and (iii) hold.*

In fact, (i) is not really essential, if we are willing to introduce an equivalence relation.

9. Distributive law; disjoint elements

The following material belongs logically directly after §3, and is independent of the results of §§4–8 above.

THEOREM 13. *In any l -group, we have for all a, b ,*

$$(18) \quad a - (a \cap b) + b = b \cup a.$$

PROOF. Substituting a for x and b for y in formula (3), Theorem 6, we get (18) explicitly.

COROLLARY 1 (Dedekind¹³). *In any commutative l -group,*

$$(19) \quad a + b = (a \cap b) + (a \cup b) \quad \text{for all } a, b.$$

In Example 3, the *modular law* (19) specializes to the celebrated identity $ab = (a, b) [a, b]$ of number theory. It also specializes, setting $b = 0$, to

COROLLARY 2. *For any a , $a = a^+ + a^-$.*

In words, each element a is the sum of its positive part and its negative part (so-called *Jordan decomposition*).

THEOREM 14. *Any l -group is a distributive lattice¹⁴.*

¹¹ This result was communicated orally to the author.

¹² This result seems to have been known for ideals, but not in abstracto. The author has been unable to find a precise reference; cf. Krull's "Idealtheorie."

¹³ Discovered in 1897; cf. Ges. Werke, Brunswick, 1931, vol. II, p. 133, formula (13); rediscovered by H. Freudenthal, "Teilweise geordnete Moduln," Amsterdam Proc. 39 (1936), p. 642.

¹⁴ In the commutative case, discovered by Dedekind, *op. cit.*, p. 135, formulas (18)–(19); rediscovered by Freudenthal, *op. cit.* p. 642, formulas (3.2).

PROOF. Bergmann has shown ("Lattice Theory", p. 75) that a lattice is distributive if and only if $a \wedge x = a \wedge y$ and $a \vee x = a \vee y$ imply $x = y$. But they imply by (18),

$$x = (a \wedge x) - a + (x \vee a) = (a \wedge y) - a + (y \vee a) = y.$$

THEOREM 15. In any l -group¹⁵, we have

$$(20') \quad a \wedge b = 0 \quad \text{and} \quad a \wedge c = 0 \quad \text{imply} \quad a \wedge (b + c) = 0,$$

$$(20'') \quad a \vee b = 0 \quad \text{and} \quad a \vee c = 0 \quad \text{imply} \quad a \vee (b + c) = 0.$$

FIRST PROOF. By hypothesis and formula (1'), $c = (a \wedge b) + c = (a + c) \wedge (b + c)$. Substituting,

$$0 = a \wedge c = a \wedge (a + c) \wedge (b + c) = a \wedge (b + c),$$

since $a \leq a + c$. The second conclusion follows by duality.

SECOND PROOF. Since a, b, c are positive, clearly $a \wedge (b + c) \geq 0$. But by the distributive law (1'),

$$\begin{aligned} 0 &= 0 + 0 = (a \wedge b) + (a \wedge c) \\ &= a + a \wedge a + c \wedge b + a \wedge b + c \geq a \wedge (b + c), \end{aligned}$$

proving (20'). Formula (20'') follows dually.

We can reword Theorem 15 in terms of the important concept of disjointness.

DEFINITION. Two positive elements a and b will be called disjoint—in symbols, $a \perp b$,—if and only if $a \wedge b = 0$.

In Example 3, this specializes to the concept of relative primeness. Theorem 15 asserts that the set of positive elements disjoint to any a is closed under addition. Furthermore, if in Theorem 13 we assume $a \wedge b = 0$ and apply the commutative law to $b \vee a$, we get the

LEMMA 1. Disjoint (positive) elements are permutable,

$$(21) \quad \text{If } a \wedge b = 0, \text{ then } a + b = b + a.$$

LEMMA 2. If $b \wedge c = 0$, then $(b - c)^+ = b$ and $(b - c)^- = -c$.

PROOF. By our preceding formulas, $(b - c) \vee 0 = (b \vee c) - c = b - (b \wedge c) + c - c = b$, and dually.

LEMMA 3. If $na \geq 0$, then $a \geq 0$.

PROOF. Expanding by the distributive law (1'), $n(a \wedge 0) = na \wedge (n - 1)a \wedge (n - 2)a \wedge \cdots \wedge a \wedge 0$. But if $na \wedge 0 = 0$, this equals $(n - 1)a \wedge (n - 2)a \wedge \cdots \wedge a \wedge 0 = (n - 1)(a \wedge 0)$. Now cancelling, we get $a \wedge 0 = 0$, as desired.

¹⁵ In the commutative case, observed by Dedekind, *op. cit.*, p. 132; Proof 1 is Dedekind's, Proof 2 is von Neumann's. Observe that in the proof, no restriction need be put on the group operation (e.g., associativity); only distributivity is needed. Theorem 15 can be generalized (§§27, 28).

Combining Lemma 3 with its dual, we get

THEOREM 16. *In an l-group, every element is of infinite order except the identity.*

Another corollary is the fact that, in any commutative l-group, $na \geq nb$ implies $n(a - b) \geq 0$, and so $a \geq b$. The author has been unable to prove the plausible conjecture that this remains true in any l-group.

LEMMA 4. *The positive and negative parts of any element are disjoint; in symbols,*

$$(22) \text{ For any } a, (a \cup 0) \wedge (-a \cup 0) = a^+ \wedge (-a^-) = 0.$$

PROOF. Clearly $-(a \wedge 0) = (-a \cup 0)$; hence the two left-hand terms are equal. But now by the distributive law, $(a \cup 0) \wedge (-a \cup 0) = (a \wedge -a) \cup 0$, so we need only show that $-(a \wedge -a) = -a \cup a \geq 0$. But clearly $a \cup -a \geq a \wedge -a$; hence, subtracting, $(a \cup -a) - (a \wedge -a) = (a \cup -a) - (-(a \cup -a)) \geq 0$, or $2(a \cup -a) \geq 0$. Now use Lemma 3 with $n = 2$.

10. Free l-groups; absolute

Now let a be any element of any l-group, and set $b = a \cup 0$, $c = -a \cup 0$, so that b and c are positive and disjoint, and $a = b - c$ (cf. Cor. 2 of Thm. 13 and Lemma 4 above). Further, by Theorem 15 and induction, $b \perp nc$ and $mb \perp nc$ for all positive integers m and n . Further, by Lemma 1, b and c are permutable, and so generate a commutative group, in which, for all integers m and n ,

$$(mb + nc) \pm (m'b + n'c) = (m \pm m')b + (n \pm n')c.$$

Finally, $(mb + nc)^+$ is $mb + nc$ unless m or n is negative, is 0 if m and n are negative, is (by Lemma 2 above and the disjointness of positive integral multiples of b and c) mb if n is negative but m is not, and is nc if m is negative but n is not.

It follows that the $mb + nc$ form an l-subgroup, which is closed under lattice and group operations (Theorem 8), and is homomorphic with the l-group of all couples (m, n) of integers, in which $(m, n) \geq 0$ means that $m \geq 0$ and $n \geq 0$. We shall (cf. §16) refer to this as the *square* of the l-group of the integers under addition.

THEOREM 17. *The free l-group with one generator is isomorphic with the square of the l-group of integers under addition.*

In this group, a appears as the element $(1, -1)$, a^+ as $(1, 0)$, and a^- as $(0, -1)$. We can read off various corollaries from this representation.

THEOREM 18. *In any commutative l-group A , the correspondence $a \rightarrow na$ is, for any positive integer n , an isomorphism of A onto an l-subgroup of itself.*

PROOF. By pure group theory, it is a group homomorphism; by Theorem 16, it is a group isomorphism; by Theorem 17, we get $(na)^+ = (n, -n)^+ = (n, 0) = na^+$, and so it is isomorphic with respect to the unary operation of taking the positive part; by formulas (4)–(5), it is therefore a lattice isomorphism.

DEFINITION. By the absolute $|a|$ of an element a of an l-group, is meant $a \cup -a$.

THEOREM 19. *In any l-group, we have identically:*

- (23) *If $a \neq 0$, then $|a| > 0$, while $|0| = 0$*
 (24) *$|na| = |n| \cdot |a|$ for any integer n ,*
 (25) *$|a| = a^+ - a^-$,*
 (26) *$|a - b| = (a \cup b) - (a \cap b)$,*
 (27) *$|(a \cup b) - (a^* \cup b)| \leq |a - a^*|$ and dually.*

PROOF. Formulas (23)–(25) are special cases of the representation of Theorem 17. Again, using (25),

$$|a - b| = ((a - b) \cup 0) - ((a - b) \cap 0) = ((a \cup b) - b) - ((a \cap b) - b)$$

from which (26) follows by group algebra. Finally, to prove (27), expand the left-hand side by (26) to get $a \cup b \cup a^* - (a \cup b) \cap (a^* \cup b)$, whence by the distributive law, $|(a \cup b) - (a^* \cup b)| = (a \cup a^*) \cup b - (a \cap a^*) \cup b$. This reduces (27) to the case $a \geq a^*$, or $a = a^* + t$ ($t \geq 0$). But $((a^* + t) \cup b) = a^* \cup (b - t) + t \leq (a^* \cup b) + t$, which takes care of this special case.

REMARK. In a commutative l-group, we can also prove the triangle inequality $|a + b| \leq |a| + |b|$, but this does not seem to hold in general; also, the author has been unable to generalize Theorem 7.8 of "Lattice Theory" to l-groups which are not commutative.

Concerning the free l-group with two or more generators, much less can be said. As an Abelian group, one can show that it has an infinite number of disjoint independent elements. On the other hand, using the three distributive laws (1), (1'), and that of Theorem 14, one can represent every element as a finite meet of finite joins of finite sums

$$\bigwedge_i \bigvee_j \sum n_k^{i,j} a_k$$

of the given generators a_k and their inverses.¹⁶

11. l-ideals

It is well-known that the different homomorphic images of a given abstract algebra can all be found by enumerating its different congruence relations.¹⁷ Also, with any group, the congruence relations correspond one-one with normal subgroups: to each normal subgroup N of a group G corresponds the congruence relation dividing G into the cosets of N . Therefore, the congruence relations

¹⁶ The construction is identical with that used to prove Theorem 5.13 of "Lattice Theory."

¹⁷ By a "congruence relation" on an abstract algebra with binary operations is meant an equivalence relation (i.e., reflexive, symmetric and transitive relation) denoted \equiv which has, if \cdot is any binary operation, the "substitution property": (S) $a \equiv a'$ implies $a \cdot b \equiv a' \cdot b$ and $b \cdot a \equiv b \cdot a'$.

on an l -group are those decompositions into cosets of normal subgroups which have the substitution property (S) for the two lattice operations—or equivalently, by (4)–(5), make $a \equiv b$ imply $a^+ \equiv b^+$. But these are easy to describe.

DEFINITION. By an l -ideal of an l -group G , is meant a normal subgroup of G which contains with any a , also all¹⁸ x with $|x| \leq |a|$.

Clearly G and 0 are l -ideals of G ; they are called *improper* l -ideals; all other l -ideals of G are called *proper* l -ideals.

It is a corollary that any l -ideal is a *convex* l -subgroup in the sense of containing with any a and b , also $-a$, $a + b$, $a \wedge b$, $a \vee b$, and every x between $a \wedge b$ and $a \vee b$. Indeed, if $a \wedge b \leq x \leq a \vee b$, then

$$\begin{aligned} |x| &= x \vee -x \leq (a \vee b) \vee -(a \wedge b) \\ &= a \vee b \vee -b \vee -a = |a| \vee |b| \leq |a| + |b|. \end{aligned}$$

THEOREM 20. The congruence relations on any l -group A are the partitions of A into the cosets of its different l -ideals.

PROOF. If N is the set of elements congruent to 0 under a congruence relation, then $a \in N$ and $|x| \leq |a|$ imply $a \wedge -a \leq x \leq a \vee -a$; hence $0 \wedge 0 \leq x \leq 0 \vee 0 \bmod N$, and so $x \in N$. Conversely, if N is an l -ideal, then $x \equiv x' \bmod N$ implies $|(x \vee y) - (x' \vee y)| \leq |x - x'|$ by (27), and therefore $x \vee y \equiv x' \vee y \bmod N$. Using left-right symmetry and duality, we see that N defines a congruence relation with respect to both lattice operations, completing the proof.

LEMMA 1. If $x \leq a + b$, where x, a, b , are positive, then $x = s + t$, where $0 \leq s \leq a$, $0 \leq t \leq b$.

PROOF. Set $t = x \wedge b$; then $x = s + t$, where $0 \leq s \leq x - (x \wedge b) = x \vee b - b \leq (a + b) - b = a$ and $0 \leq t \leq b$, as desired.

THEOREM 21. The l -ideals and any l -group form a complete distributive sublattice of the (modular) lattice of all its normal subgroups.

PROOF. Clearly, any intersection of l -ideals is itself an l -ideal. To prove that the sum $S + T$ of any two¹⁹ l -ideals S and T is an l -ideal, suppose that $s \in S$, $t \in T$, and $|x| \leq s + t$. Then for some $s' \in S$, $t' \in T$, $-(s + t) = s' + t'$, and so

$$|s + t| = (s + t) \vee (s' + t') \leq (0 \vee s \vee s') + (0 \vee t \vee t').$$

Hence $x^+ \leq |x| \leq |s + t| \leq s'' + t''$ ($s'' \in S$, $t'' \in T$). Using Lemma 1, we can show now that $S + T$ contains x^+ , likewise $-x^-$, and so $x = x^+ + x^-$. Therefore $S + T$ is an l -ideal.

¹⁸ The terminology is that of Stone (*op. cit.*); the concept is due to the author, who called l -ideals "normal subspaces"; F. Riesz, "Sur la théorie générale des opérations linéaires", *Annals of Math.* 41 (1940), pp. 174–206, called them "Families presque complètes." Another good term would be "absolute (normal) subgroup." Kakutani uses l -ideals in a slightly different sense.

¹⁹ From this it follows that the sum of any number of l -ideals is an l -ideal—by the general logical principle that for any "closure" involving only finite operations, the closure of any family of "closed" sets is the set-union of joins of finite subfamilies of "closed" sets.

It remains to prove that if S , T , and U are l-ideals, then $S \cap (T + U) = (S \cap T) + (S \cap U)$. But since, in any case, $S \cap (T + U) \supseteq (S \cap T) + (S \cap U)$ by the lattice-theoretic semi-distributive law, and $x = x^+ - (-x^-)$, it suffices to show that every positive x in $S \cap (T + U)$ is in $(S \cap T) + (S \cap U)$. But $x \in S \cap (T + U)$ means that $x \in S$ and $x = t + u$ ($t \in T$, $u \in U$). Hence, as above, $x = |t + u| \leq t'' + u''$, where $t'' \in T$, $u'' \in U$ are positive. Therefore, by Lemma 1, $x = t' + u'$, where $t' = x \cap t''$ is in $S \cap T$ and $u' \leq x \cap u''$ is in $S \cap U$. This proves $x \in (S \cap T) + (S \cap U)$, as desired.

COROLLARY. *The congruence relations on any l-group form a complete distributive lattice.*²⁰

REMARK. If A is any commutative l-group, and T is any l-ideal of an l-ideal S of A , then T is itself an l-ideal of A : the property of being an l-ideal is thus hereditary. This follows because any subgroup of a subgroup is itself a subgroup, and by the transitivity of inclusion. However, as Example 9 illustrates, the same law does not hold for all non-commutative l-groups—essentially because a normal subgroup of a normal subgroup of a group G need not be normal in G .

12. Disjoint l-ideals

The following sections, through §20, will deal with non-commutative l-groups only incidentally. In the main, they will be devoted to obtaining a more complete picture of the structure of commutative l-groups, including a determination of all possible structure lattices of finite length, and of all those "simple" commutative l-groups which have no proper l-ideals.

DEFINITION. Two elements a and b of an l-group G are called disjoint if and only if $|a| \cap |b| = 0$.

THEOREM 22. *The set $\{a\}^*$ of all elements disjoint from any fixed element a is a subgroup which contains with any b , all x satisfying $|x| \leq |b|$.*

PROOF. By (23), $\{a\}^*$ contains 0; by Theorem 15, it is closed under addition; since $|-b| = |b|$, it contains with any element its group inverse; hence it is a subgroup. The second assertion follows by the monotonicity law.

COROLLARY. *In a commutative l-group, the set of all elements disjoint from any fixed element is an l-ideal.*

Example 9 shows that, in the non-commutative case, the set need not be a normal subgroup.

We note also, since $\{a\}^*$ cannot contain a unless $a = 0$, either $a = 0$, $\{a\}^* = 0$, or $\{a\}^*$ is a proper l-ideal. This suggests the concept of a weak unit.

DEFINITION. An element a of an l-group is called a weak unit²¹ if the only element disjoint to it is 0.

²⁰ This is the "structure lattice" of the l-group in the sense of the author, "On the structure of abstract algebras," Proc. Camb. Phil. Soc. 31 (1935), p. 450. It describes the structure (in the usual sense) of the l-group. Theorems 20–21 are due to the author.

²¹ The concept is due to Freudenthal, op. cit.; the useful terms "weak unit" and "strong unit" (infra) to Bohnenblust. We note that any separable Banach lattice has a weak unit.

13. Simply ordered groups

A partially ordered set is called "simply ordered" when, of any two elements, one includes the other, so that

P_4 . Given x, y , either $x \geq y$ or $y \geq x$.

This automatically implies $L'-L''$.

DEFINITION. A simply ordered group²² is an l-group in which P_4 holds.

We note without proof the following trivial results. An l-group is simply ordered if and only if, for any a , either a or its inverse $-a$ is positive. An l-group is simply ordered if and only if every subgroup is an l-subgroup. The structure lattice of any simply ordered l-group is itself simply ordered (a chain).²³

DEFINITION. Two l-ideals of an l-group are called disjoint if and only if their intersection is 0.

It is easy to show that this is the case if and only if every element of the first ideal is disjoint from every element of the second.

THEOREM 23. A commutative l-group has two disjoint proper l-ideals unless it is simply ordered.

PROOF. Unless the l-group is simply ordered, it has an element a which is neither positive nor negative, so that neither a^+ nor a^- is 0. Hence $S = \{a^+\}^*$ will be a proper l-ideal containing a^- but not a^+ . Moreover the set S^* of all elements disjoint from all elements of S will contain a^+ but not a^- . Furthermore, being an intersection of l-ideals, it will be an l-ideal. Finally, every element of S is disjoint from every element of S^* .

COROLLARY. The structure lattice of a commutative l-group A is simply ordered if and only if A is simply ordered.

Example 9 shows that the hypothesis of commutativity is essential in the preceding results.

DIGRESSION. We have seen (Theorem 16) that in an l-group, every element is of infinite order. We shall now show that, in the commutative case, this is the only group-theoretic restriction implied by being an l-group.

THEOREM 24 (F. Levi²⁴). Any abstract commutative group whose elements are all of infinite order, is the additive group of a simply ordered l-group.

PROOF. Let A be any Abelian group without any element of finite order except the identity. By a well-ordered rational basis for A , we mean a well-

In fact, if $\{x_i\}$ is any everywhere dense countable set of positive elements, and $\lambda_i = 1/2^i \parallel x_i \parallel$ for all i , then $e = \sum \lambda_i x_i$ is a weak unit.

²² Often called an "ordered group"; this is consistent with the terminology "semi-ordered group" for what we have called a "partially ordered group."

²³ For if the l-ideal S contains an element not in the l-ideal T , then the absolute of this element must exceed (not being included in) the absolute of every element of T , so that $T \leq S$.

²⁴ "Arithmetische Gesetze im Gebiete diskreter Gruppen," Rendic. Palermo 35 (1913), pp. 225-236.

ordered (finite or infinite) subset of elements a_α of A such that every non-zero element of A is a finite rational combination $n_1 a_{\alpha(1)} + \cdots + n_r a_{\alpha(r)}$ ($\alpha(1) < \cdots < \alpha(r)$) of the a_α , while $\sum n_i a_{\alpha(i)} = 0$ implies every $n_i = 0$ —or equivalently, $\sum (m_i/n_i) a_{\alpha(i)} = 0$ implies that every $(m_i/n_i) = 0$. The existence of a well-ordered rational basis can be proved directly by transfinite induction, just as in the case of vector spaces.

Moreover relative to such a basis, any element of A not the identity may be called positive or negative according as its first non-zero coefficient is positive or negative. This "lexicographic" ordering of A clearly defines from it a simply ordered group (commutative l-group).

COROLLARY. *A commutative group is the additive group of an l-group if and only if it is without elements of finite order except the identity.*

14. Archimedean l-groups

A gross way of comparing the magnitude of elements of l-groups is given by the following

DEFINITION. *An element a of an l-group is called incomparably smaller than a second element b (in symbols, $a \ll b$) if and only if $na < b$ for any integer n .*

Otherwise stated, $a \ll b$ means that b is an upper bound for the entire cyclic subgroup generated by a . Thus in Example 4, $(0, 1) \ll (1, 0)$. It is easily verified that the relation \ll is antisymmetric and transitive; it is closely related to the concept of an Archimedean l-group.

DEFINITION. *An l-group is called Archimedean if and only if $a \ll b$ implies $a = 0$. (I, II, P_1 - P_3)*

The independence of the Archimedean property just stated from the lattice property I' - L'' is illustrated by the easily proved fact that *any subgroup of an Archimedean l-group is itself Archimedean* with respect to the same order relation, whether it is an l-subgroup or not.

The Archimedean property can be formulated in other ways. It amounts to asserting that the l-group has no bounded subgroups except 0. It is equivalent to requiring that if the set of all positive multiples of a has an upper bound, then $a \leq 0$ (Clifford). In the case of l-groups, using Cor. 2 of Thm. 13, it is equivalent to the apparently weaker requirement that if $a > 0$, then the sequence $a, 2a, 3a, \cdots$ has no upper bound.

In a simply ordered group, the Archimedean property is thus equivalent to the traditional condition that for any $e > 0$ and any $b, ne > b$ for all sufficiently large positive integers n . This means that if we let U denote the set of all rational numbers m/n such that $nb \geq me$, and L the set of those such that $nb \leq me$ (n positive), we get non-void sets. Moreover L and U together include all elements (by P_4), and have at most one element in common. Hence they are the two halves of a *Dedekind cut*. Again, no two distinct elements b and b' can determine the same cut, or we would have $(b - b') \ll e$. Finally, by the monotonicity law (2), addition of elements is isomorphic to the addition of cuts. We conclude

THEOREM 25. *Any simply ordered Archimedean l-group is isomorphic to a subgroup of the additive group of all real numbers, and so is commutative.*²⁵

THEOREM 26. *An Archimedean l-group may have a non-Archimedean homomorphic image.*

PROOF. Consider the l-quotient-group of the l-group of all functions on the interval $0 \leq x < +\infty$, modulo the l-ideal of bounded functions. In this, $x^2 > 0$, yet $x^2 \ll x^4$.

DIGRESSION. We have seen that in any l-group, for any element a , the equation $nx = ma$ ($n \neq 0$) has at most one solution, which we can denote $(m/n)a$ if it exists. It is worth remarking now that in any Archimedean l-group, we can define uniquely scalar products λa of a by any real number λ . To see this, suppose a positive; there is at most one x such that $(m/n)a < x$ for all $m/n < \lambda$ and $(m/n)a > x$ for all $m/n > \lambda$. (If two, x and x' , then $x - x' \ll a$.) This x we may denote λa , and prove that, whenever all terms exist, the usual laws of the vector calculus hold.

15. Strong units: principal l-ideals

We have just seen that in any Archimedean simply ordered l-group, to any $e > 0$ and b corresponds a positive integer n such that $ne > b$. This may be generalized.

DEFINITION. *By a strong unit of an l-group A , is meant²⁶ an element $e \in A$ such that for any $b \in A$, $ne > b$ for some positive integer n .*

Thus a strong unit must be positive. Many l-groups do not have any strong unit. For example, the l-group of all continuous real functions on the domain $0 \leq x < +\infty$ has the weak unit $f(x) = 1$ but no strong unit; this is a weak corollary of the Theorem of du Bois-Reymond.²⁷ On the other hand, in the l-group of all bounded real functions on any domain, the function $f(x) = 1$ is a strong unit. We also note

LEMMA 1. *Any strong unit is a weak unit.*

PROOF. For any e , $e \wedge a = 0$ implies $ne \wedge a = 0$ for all e (Thm. 22). But if e is a strong unit, $ne > a$ for some n and so $e \wedge a = 0$ implies $a = ne \wedge a = 0$, whence e is a weak unit.

Even in l-groups without strong units, l-ideals may have strong units. In fact, in any commutative l-group, every positive element is a strong unit for an appropriate l-ideal.

THEOREM 27. (F. Riesz.²⁸) *In a commutative l-group, for any $a > 0$, the set $J(a)$ of all b such that $|b| \leq na$ for some positive integer n forms an l-ideal having a as strong unit. Moreover $J(a)$ is the smallest l-ideal which contains a .*

²⁵ This result is due to H. Cartan, "Un théorème sur les groupes ordonnés," Bull. Sci. Math. 63 (1939), 201-205.

²⁶ The concept goes back to Archimedes; the term to Bohnenblust.

²⁷ Cf. for instance, G. H. Hardy, "Orders of Infinity," Cambridge Tracts, 2d ed., 1924, p. 8. In this example, our relation $a \ll b$ is practically the usual relation $f = o(g)$.

²⁸ F. Riesz, *op. cit.*, p. 188.

PROOF. If $|b| \leq ma$ and $|c| \leq na$, then clearly $|b \pm c| \leq (m+n)a$; while if $|b| \leq ma$ and $|x| \leq |b|$, then $|x| \leq ma$; hence $J(a)$ is an l-ideal. Obviously, a is a strong unit of $J(a)$. Finally, any l-ideal containing a must contain every na and so all b with $|b| \leq na$.

COROLLARY. Any commutative non-Archimedean l-group has a proper l-ideal.

For if $a \ll b$ for some $a \neq 0$, then $J(|a|)$ is an l-ideal which fails to contain b , yet contains $a \neq 0$, and so is proper.

DEFINITION. An l-ideal of an l-group will be called principal if and only if it has a strong unit.

THEOREM 28. If the structure lattice of a commutative l-group has finite length r , every l-ideal is principal.

PROOF. Let J be an l-ideal of such an l-group A . The case $J = 0$ is trivial. If $J > 0$, choose any $a_1 \neq 0$ in J and form $J(|a_1|)$. If $J > J(|a_1|)$, choose any a_2 in J but not in $J(|a_1|)$ and form $J(|a_1| + |a_2|)$. After repeating this process at most r times, we will get a principal l-ideal equal to J .

THEOREM 29. In any commutative l-group A , the principal l-ideals form a topologically dense sublattice of the structure lattice of A .

PROOF. It can be proved easily that

$$J(a \wedge b) = J(a) \wedge J(b) \quad \text{and} \quad J(a + b) = J(a) + J(b),$$

hence they form a sublattice. This sublattice is dense in the structure lattice of A , since any l-ideal J is the supremum (in fact, set-union) of the finite joins $\vee_i J(a_i)$ of the principal l-ideals contained in J , and these form an ascending directed set of principal l-ideals which thus converges to its supremum in the sense of Moore-Smith.

16. Extension problem

It is natural to say that an l-group is *simple* if and only if it has no proper l-ideals—or, equivalently, no proper congruence relations. Analogy with pure group theory then suggests the program²⁹ of first determining all simple l-groups, and then showing how the most general l-group whose “structure lattice” is of finite length can be built up from its simple quotient-l-groups.

The first problem has been solved in the commutative case. Indeed, a simple commutative l-group must be simply ordered (by Theorem 23) and Archimedean (by the Cor. of Thm. 28). Hence (by Theorem 26) we have

THEOREM 30. The only commutative simple l-groups are the subgroups of the additive group of real numbers.

The second problem involves in particular the specific task of enumerating all the l-groups having a given l-ideal J and l-quotient-group A/J (“Extension Problem”). While not attempting a complete solution of this, some fragmentary results may be stated.

²⁹ The logical outline is the same, but the technique is very different. Cf. O. Schreier, “Über die Erweiterungen der Gruppen,” Monats. Math. u. Phys. 34 (1926), p. 165, and Hamb. Abh. 4 (1927), pp. 321–346.

Given two l-groups S and T , one can form the l-group ST of all couples (s, t) ($s \in S, t \in T$), where both the group operation and the lattice operations are performed on the S -components and T -components independently, so that

$$(s, t) \circ (s', t') = (s \circ s', t \circ t')$$

where \circ is $+$, \wedge , or \vee . This is the direct union of S and T in the sense of universal algebra; we shall call it the *cardinal product* ST of S and T . The elements $(0, t)$ form an l-ideal of $ST = A$ isomorphic with T , and the l-quotient-group A/T is isomorphic with S . Hence *the extension problem always has at least one solution.*

One can also form the lexicographic or *ordinal product*³⁰ $S \circ T$ of any two l-groups. This consists of the couples (s, t) ($s \in S, t \in T$) just as before. But the set of positive elements consists of those couples (s, t) with $s > 0$ or $s = 0$ and $t \geq 0$, instead of those with $s \geq 0$ and $t \geq 0$ as in the case of cardinal products. In any case, $S \circ T$ is a partially ordered group. If S is *simply* ordered, it is an l-group, in which the elements $(0, t)$ form as before an l-ideal isomorphic with T , whose l-quotient-group is isomorphic with S . Hence if S is simply ordered, the extension problem has at least two solutions.

17. Direct decompositions

In the cardinal product $ST = A$, both S and T correspond to l-ideals. Moreover they correspond to *complementary* l-ideals, in the usual sense that $S \wedge T = 0$ and $S + T = A$. Just as in the case of pure group theory, the converse also holds.

THEOREM 31. *An l-group A is isomorphic to the cardinal product ST if and only if it contains complementary l-ideals isomorphic with S and T respectively.*

PROOF OF CONVERSE.³¹ Suppose A has l-ideals S and T . Then by group theory, each element $a \in A$ has a unique representation $a = s + t$ ($s \in S, t \in T$), while group operations are performed on the S - and T -components independently. As regards order, $s + t \leq s' + t'$ if and only if

$$(-s' + s) \leq (t' - t) \leq |t' - t|,$$

whence $(-s' + s) \leq |t' - t| \wedge |-s + s'| \in T \wedge S = 0$, and likewise $t - t' \leq 0$. This means $s \leq s'$ and $t \leq t'$, q.e.d.

From the preceding result, Theorem 21, and the general theory of distributive lattices, we obtain just as in "Lattice Theory," Theorem 5.15, the following corollaries.

THEOREM 32. *Any two representations of an l-group as a cardinal product have a common refinement.*

³⁰ For the general significance of cardinal and ordinal products, cf. the author's article "Generalized arithmetic," to appear in the Duke Journal of Mathematics. It is shown there that the ordinal product of two lattices is itself a lattice if and only if the left-factor is simply ordered, or the right-factor has universal bounds.

³¹ For a brief proof, relying more heavily on principles of universal algebra, cf. also "Lattice Theory," p. 110, below Theorem 7.11.

COROLLARY. *If the structure lattice of an l -group A has finite length, then A has a unique representation as the cardinal product of indecomposable factors.*

18. Main structure theorem

In the present section, we shall show that the structure of a commutative l -group is of a very special kind. Indeed, by Theorem 23, any commutative l -group in which the l -ideal 0 is meet-irreducible (or "prime"), is simply ordered. From this (cf. footnote 23) we conclude

LEMMA 1. *The structure lattice of a commutative l -group in which the l -ideal 0 is meet-irreducible, is a chain.*

But now if J is any l -ideal of an l -group A , the l -ideals of A which contain J form a lattice isomorphic with the structure lattice of A/J ; indeed, this is a principle of universal algebra, holding for all congruence relations. Combining this result with Lemma 1, we get

LEMMA 2. *The elements of the structure lattice of any commutative l -group which contain any meet-irreducible element, form a chain (simply ordered set).*

If we apply Lemma 2 to the general representation theory of finite distributive lattices ("Lattice Theory," Theorem 5.3), we get a conclusive result.

Any distributive lattice L of finite length may be described in terms of the partially ordered set X of its meet-irreducible elements a_i . Every element $c \in L$ is the meet $\bigwedge a_i$ of the set S_c of the meet-irreducible elements which contain c . Moreover, as in "Lattice Theory," Theorem 5.3, the correspondence $c \rightarrow S_c$ is a dual isomorphism between L and the " J -closed" subsets of X —i.e., the subsets of X which contain with any a_i all $a_j \geq a_i$.

This clearly applies to the structure lattice of any l -group, provided it has finite length. Moreover if the l -group is commutative, Lemma 2 restricts X greatly.

DEFINITION. *A partially ordered system X is called a semitree if, for any element $a \in X$, the set of all $x \leq a$ is a chain; it is a tree if it has a least element 0 . The dual of a tree (semitree) is called a root (semiroot).³²*

We have shown that the meet-irreducible ("prime") l -ideals of any l -group form a semiroot. But now it is easy to show that any finite semiroot is the sum of the subsets contained in its different maximal elements: the elements underneath its different maximal elements form components having no connection with each other (no common subelements or superelements).

Consequently, either the structure lattice contains complemented elements (namely, the meets of the sets of elements under the different maximal meet-irreducible elements), or the set X of meet-irreducible elements has a I . In the first case, the l -group is directly decomposable, by Theorem 31. In the second case, the J -closed subset consisting of I alone is a least non-void J -closed subset,

³² The Hasse diagram of any "tree" looks like a tree, and that of a "root" like a root (tree upside down). Further, the graph of a "tree" (or root!) is a tree in the technical sense of the theory of graphs. G. Kurepa has studied roots extensively, under the name of "tableaux ramifiés."

which thus corresponds under our dual isomorphism to a *greatest* proper l -ideal. We can state our result as follows.

THEOREM 33. *A commutative l -group whose structure lattice has finite length, either (i) is a cardinal product, or (ii) has a unique maximal proper l -ideal.*

19. Solution of extension problem

We shall now show that a commutative l -group A with a unique maximal proper l -ideal J is a kind of mixed ordinal product of A/J and J . This will give us a method for constructing, by successive extensions, all commutative l -groups having finite structure lattices.

First, an element a of A not in J must be either positive or negative. For consider the sum of the l -ideals generated by a^+ and a^- ; it contains a , hence is not contained in J , hence it is A . But this expresses A as a sum of disjoint l -ideals; by hypothesis, A is join-irreducible; hence one of the l -ideals is A and the other (being disjoint) is 0 , and a^+ or a^- is 0 , as desired.

Second, a is positive or negative in A according as it is positive or negative in A/J , since a homomorphism carries positive elements into positive elements and dually. Hence A is determined to within isomorphism by its group structure, the order structure of J , and the order structure of A/J . The positive elements of A are those which have their (A/J) -component greater than zero, or have their (A/J) -component equal to zero and their J -component positive.

This definition gives, conversely, from any abstract Abelian group A which has a lattice-ordered subgroup J and simply ordered quotient-group A/J , an l -group which may be called a *mixed ordinal product* of J and A/J . Clearly the mixed ordinal products of J and A/J correspond one-one to the solution of the group-theoretic extension problem of finding all Abelian groups A with a subgroup isomorphic with J and a quotient-group isomorphic with A/J . In case A is the direct union of J and A/J , we get the pure ordinal product; otherwise, we get something different.

Now by Theorem 33, and induction (cf. the last Remark of §11) on the length of the structure lattice, we get

THEOREM 34. *Any commutative l -group whose structure lattice has finite length can be built up from simple l -groups by forming successive cardinal products and mixed ordinal products.*

This result can be applied directly to vector lattices. It is known³³ that the additive group of real numbers is the only simple vector lattice. Moreover it can be shown that for finite-dimensional vector lattices, the only group-theoretic solution of the extension problem is given by the direct union. We conclude

COROLLARY 1. *Any vector lattice of finite dimension can be built up from the group of real numbers under addition by repeated formation of cardinal and ordinal products.*

³³ Mr. Murray Mannos, a graduate student at Harvard University, is writing a dissertation on vector lattices of finite dimension which includes this and many other results.

We can state this somewhat cabalistically, using the generalized arithmetic notation of the author, as

COROLLARY 2. *The most general vector lattice of finite dimension is ${}^Y R \#$, where $R \#$ denotes the additive group of real numbers, and Y denotes the most general semiroot.³⁴*

Incidentally, the structure lattice of ${}^Y R \#$ is $B^{Y'}$, where Y' denotes the semi-tree dual to Y , ordinal exponentiation is replaced by cardinal exponentiation, and B is the chain of two elements.

Going back to Lemma 2 of §18, the discussion of distributive lattices following it, and using Corollary 2 for the converse, we get a final result.

THEOREM 35. *A lattice of finite length is the structure lattice of a commutative l-group, if and only if it can be written B^Y , where Y is the most general semitree.*

20. Subdirect decompositions

We shall now consider the representations of commutative l-groups as l-subgroups of cardinal products of smaller l-groups—or, as we shall say for short, as *subdirect* products.

Just as in the case of groups (cf. "Lattice Theory," p. 52) it may be shown that the representations of an l-group as a subdirect product correspond one-one to choices of sets of l-ideals having 0 for meet. In the case of structure lattices of finite length, we can thus show that commutative l-groups are subdirect products of l-groups in which 0 is meet-irreducible, and hence (§18, Lemma 1) of simply ordered l-groups. We shall now show that the restriction to the case of structure lattices of finite length is unnecessary.

LEMMA 1. *Let a be any non-zero element of a commutative l-group A . There exists an l-ideal J in A such that $a \notin J$ yet J/J is meet-irreducible in A/J .*

PROOF. By transfinite induction, we can construct a maximal l-ideal J which does not³⁵ contain a . It follows that any l-ideal of A which properly contains J will contain $J(a)$; hence J is meet-irreducible. We infer that the meet of all meet-irreducible l-ideals of A is 0 in any case; hence that A is a subdirect product of l-quotient-groups A/J in which 0 is meet-irreducible, and so which are simply ordered.

THEOREM 36. *Any commutative l-group is isomorphic with an l-subgroup of a cardinal product of simply ordered l-groups.³⁶*

³⁴ It has been pointed out to the author by A. H. Clifford and I. Kaplansky that Theorem 34 and its corollaries may be looked on as generalizing to the lattice-ordered case, the basic results of H. Hahn ("Über die nichtarchimedischen Grössensysteme," S.-B. Wiener Akad. Math.-Nat. Klasse Abt. IIa, 116 (1907), pp. 601–653) on the classification of simply ordered groups.

³⁵ The construction is identical with that used by Stone in constructing prime ideals in Boolean rings; it has been used so often that it will not be repeated here. Since an l-ideal J is meet-irreducible if and only if $|a| \frown |b| \in J$ implies $|a| \in J$ or $|b| \in J$, there is justification for calling the meet-irreducible l-ideals *prime* l-ideals.

³⁶ This is closely related to Satz 14 of P. Lorenzen, "Abstrakte Begründung der multiplikativen Idealtheorie," Math. Zeits. 45 (1939), pp. 533–553.

A special problem is that of trying to make the simply ordered l-groups *Archimedean*, so as to get a representation by means of real functions. For this to be possible, the original l-group must certainly be Archimedean; this condition would also be sufficient if it were not for Theorem 26. Much work has been done in attacking special cases of the problem.³⁷

21. Effect of chain condition

We shall now turn our attention to l-groups in which all bounded sets have l.u.b. and g.l.b. A special case is furnished by l-groups which satisfy the chain condition.

DEFINITION. An l-group will be said to satisfy the chain condition³⁸ if and only if

(C) every non-void set of positive elements includes a minimal member.
Any element which covers 0 will be called a prime.

LEMMA 1. Any two primes are permutable.

This is a corollary of Lemma 1 of §9. It is a corollary that the primes generate an Abelian subgroup, consisting of all elements which can be expressed as sums $n_1p_1 + \cdots + n_sp_s$ of a finite number of distinct primes.

Now let $a > 0$ be given, and consider all those differences $a - \sum n_i p_i$ which are positive. By the chain condition, one of these must be minimal, and so cannot contain any prime q (otherwise $a - (\sum n_i p_i + q)$ would be smaller). Again by the chain condition, every positive element b except 0 contains a prime, namely, some minimal x such that $0 < x \leq b$. Hence our minimal difference must be 0, so that $a = \sum n_i p_i$.

But every element can be expressed as a difference of positive elements: $c = c^+ - (-c)^+$ for all c ; hence

LEMMA 2. Any element not 0 can be expressed as a sum of integral multiples of a finite number of distinct primes, as $a = n_1p_1 + \cdots + n_sp_s$.

Putting Lemmas 1-2 together, we infer that our l-group is commutative. Now if we distinguish positive and negative coefficients, we get an expression for any $a \neq 0$ as

$$a = m_1p_1 + \cdots + m_rp_r - n_1q_1 - \cdots - n_sq_s. \quad (m_i, n_j > 0)$$

Clearly a cannot be positive unless $q_j \leq m_1p_1 + \cdots + m_rp_r$ for all j . But since distinct primes are disjoint, by Theorem 15 q_j is disjoint from $\sum m_i p_i$; hence a cannot be positive unless no negative coefficients occur.

LEMMA 3. In Lemma 2, a is positive if and only if every n_i is positive.

³⁷ Cf. F. Bohnenblust, *op. cit.*; S. Kakutani, "Weak topology, bicomact set, and the principle of duality," *Proc. Imp. Acad. Tokyo* 16 (1940), pp. 63-67, Thm. 6; Stone, *op. cit.*; M. and S. Krein, *Doklady* 27 (1940), pp. 427-430; and K. Yosida, "On vector lattice with a unit," *Proc. Imp. Acad. Tokyo* 17 (1941), pp. 121-124.

³⁸ Ore uses the word "Archimedean" to mean the same thing, but our terminology is more common. In the simply ordered case, (C) implies that every non-void set of positive elements has a least member (well-ordering condition), so that the integers form the only simply ordered l-group satisfying the chain condition.

It is a corollary that a is zero (positive and negative) if and only if every n_i is positive and negative, which is absurd. It is a corollary that the representation of Lemma 1 is unique; for if a had two different representations, their formal difference would give a representation of 0. We can summarize.

THEOREM 37. *Let A be any l -group which satisfies the chain condition. Then A is commutative, and each non-zero element of A can be expressed uniquely as a sum of integral multiples of distinct primes.³⁹ Such a sum is positive if and only if no coefficient is negative.*

It is a corollary that A is determined to within isomorphism by the cardinal number of the set of its primes.

22. Application to ideal theory

This suggests an approach to the so-called "fundamental theorem of ideal theory" quite different from the modern approach,⁴⁰ and much nearer to the classical one. Let F be any field, and let H be any subring of "integers" of F which contains unity. By an *ideal* in F , we mean a subset which contains with any two elements their sum and difference, and with any element all its integral multiples. Multiplication of ideals is according to the usual definition.

It is clear that the non-zero ideals form a lattice with respect to set-inclusion, which in many important cases can be proved by extremely general arguments to satisfy the chain condition.⁴¹

It is also clear that multiplication of ideals is commutative and associative, and that ideal multiplication is distributive on addition (the lattice-join). Therefore we have all of the postulates of Theorem 7 satisfied except the existence of inverses.

It follows that, in the most important cases, in order to establish the unique factorization of ideals into primes, we need only supplement general arguments by proving that *every ideal has an ideal inverse*—or equivalently, that the product of every ideal by a suitable ideal gives a principal ideal.

23. Completeness

Many important l -groups are complete, in the sense of the following

DEFINITION. *An l -group A is called complete (σ -complete) if and only if every non-void (resp. countable) bounded set has a g.l.b. and a l.u.b.*

REMARK 1. By Theorem 2, the existence of g.l.b. implies that of l.u.b.; and

³⁹ In the commutative case, this result is essentially well-known. Cf. for example M. Ward, "Residuated distributive lattices," *Duke Jour.* 6 (1940), pp. 641-651; also A. Clifford, "Arithmetic and ideal theory of abstract multiplication," *Bull. Am. Math. Soc.* 40 (1934), p. 329, Thm. 2.

⁴⁰ For the modern treatment of E. Noether, cf. van der Waerden's "Moderne Algebra," 1st ed., vol. Z, pp. 98-102. For the classical treatment cf. D. Hilbert, "Théorie des corps de nombres algébriques," Paris, 1913. Remarks much like ours are made on p. 13 of Krull's "Idealtheorie."

⁴¹ Cf. van der Waerden, *op. cit.*, §80. By a "positive" ideal, we mean one which contains H , which is an identity for multiplication. The "negative" ideals are thus the ideals which are *integral*, in the usual terminology.

using Theorem 1, one can even show that it is enough to require that every non-void set of *positive* elements have a g.l.b.

REMARK 2. The chain condition implies completeness. For if S is a non-void set of positive elements s_α , then the *finite* meets $\bigvee s_{\alpha(i)}$ include a minimal member a by the chain condition. But every $s_\alpha \wedge a$, being itself a finite meet and so not properly contained in a , will be a . Thus a is a lower bound for S ; it obviously contains every lower bound.

Next, let x be any element of any l-group A . If s is any upper bound for the set $\{nx\}$, then by Theorem 1 so are $s + x$ and $s - x$. It follows that $\{nx\}$ cannot have a *least* upper bound unless $x \geq 0$ and $x \leq 0$.

LEMMA 1. *The set of all integral multiples of a non-zero element cannot have a l.u.b. (I, II, P₂)*

COROLLARY. *Unless $A = 0$, any l-group A contains a countable set without a least upper bound.*

It also follows that, if A is σ -complete, the set nx cannot have an upper bound (or it would have a l.u.b.). In other words,

THEOREM 38. *Any σ -complete l-group is Archimedean.*

COROLLARY. *If an l-group can be embedded group- and order-isomorphically in a complete l-group, then it is Archimedean.*

Conversely, Clifford (*op. cit.*) has proved that any commutative Archimedean l-group⁴² can be completed by cuts in the sense of Dedekind-MacNeille, to give a complete commutative l-group. Combining, we have

THEOREM 39 (Clifford). *A commutative l-group can be embedded in a complete l-group if and only if it is Archimedean. (I, II, P₁-P₃, (14))*

24. Infinite distributivity

It was proved (Theorem 1) that in an l-group, any group translation is a lattice automorphism. Consequently, it carries infinite joins and meets into infinite joins and meets, respectively. The formulas expressing this fact appear as the infinite distributive laws

$$(28) \quad \begin{aligned} a + \bigvee x_\alpha &= \bigwedge (a + x_\alpha) & a + \bigwedge x_\alpha &= \bigwedge (a + x_\alpha) \\ \bigvee x_\alpha + b &= \bigwedge (x_\alpha + b) & \bigwedge x_\alpha + b &= \bigwedge (x_\alpha + b) \end{aligned}$$

Similarly, since every correspondence of the form $x \rightarrow a - x$ is a dual automorphism, we have the formal laws

$$(29) \quad a - \bigvee x_\alpha = \bigwedge (a - x_\alpha) \quad \text{and dually.}$$

Now let $v = \bigvee x_\alpha$. Then, for all a and α ,

$$0 \leq (a \wedge v) - (a \wedge x_\alpha) \leq v - x_\alpha \quad \text{by (27).}$$

⁴² Actually, $L'-L''$ may be replaced for this purpose by the far weaker condition (14) (Moore-Smith property). In the present case, the cuts appear as so-called v -ideals; cf. Krull's v -Gruppensatz, "Idealtheorie," p. 120.

But $\Lambda(v - x_\alpha) = v - \vee x_\alpha = v - v = 0$ by (29); and by what we have just seen, $0 \leq \Lambda[(a \wedge v) - (a \wedge x_\alpha)] \leq \Lambda(v - x_\alpha)$; hence

$$0 = \Lambda[(a \wedge v) - (a \wedge x_\alpha)] = (a \wedge v) - \vee(a \wedge x_\alpha).$$

Transposing, we get the first of the further infinite distributive laws

$$(30) \quad a \wedge \vee x_\alpha = \vee(a \wedge x_\alpha) \quad \text{and} \quad a \vee \Lambda x_\alpha = \Lambda(a \vee x_\alpha);$$

the second follows by duality. Summarizing, we have

THEOREM 40 (Kantorovitch⁴³). *The infinite distributive laws (28)–(30) hold in any complete l-group.*

25. Closed l-ideals

In a complete commutative l-group, the complemented l-ideals may also be characterized in terms of closure properties. To see this, let us define for any set S of elements of an l-group G , the polar⁴⁴ S^* of S as the set of all elements disjoint from every element of S .

If S is a complemented l-ideal with complement T , then $y \in T$ implies, for all $x \in S$, that

$$||x| \wedge |y|| = |x| \wedge |y| \leq |x| \quad \text{and} \quad |y|.$$

Hence $|x| \wedge |y| \in S \cap T = 0$, and $x \perp y$, proving $y \in S^*$. Conversely, if $z = x + y$ ($x \in S$, $y \in T$) is in S^* , then

$$0 = |z| \wedge |x| = (|x| + |y|) \wedge |x| = |x|,$$

whence $z = y$ is in T . This shows $T = S^*$; by symmetry, $S = T^* = (S^*)^*$.

But now for any subset T , the set T^* is an l-ideal by Theorem 22, provided G is commutative. Further, by (30), if G is complete, then T^* is a closed l-ideal in the sense of the following definition.

DEFINITION. *An l-ideal J of a complete l-group G is called closed⁴⁵ if and only if J contains with any bounded subset $\{x_\alpha\}$, also $\vee x_\alpha$.*

REMARK. Since the correspondence $x \rightarrow -x$ leaves J setwise invariant and inverts order, it follows that J also contains Λx_α . Further, since any l-ideal is convex (§11), closure in the sense of the preceding definition is equivalent to topological closure in the intrinsic topology.⁴⁶

⁴³ "Lineare halbgeordnete Räume," Math. Sbornik, 2 (44) (1937), pp. 121–168, esp. Theorems 10–21. Kantorovitch assumed commutativity, but this does not play an essential role.

⁴⁴ It follows from the general theory of relations (cf. "Lattice Theory", §32), since the relation of disjointness is symmetric and anti-reflexive, that (i) if we denote $(S^*)^*$ by \bar{S} , then the operation $S \rightarrow \bar{S}$ is a closure operation, (ii) if we call S "closed" when $S = \bar{S}$, then any intersection of "closed" sets is itself closed, (iii) 0 is closed, (iv) the correspondence $S \rightarrow S^*$ is a dual automorphism of the lattice of "closed" sets.

⁴⁵ Closed l-ideals are the "familles completes" of F. Riesz, *op. cit.*, Riesz proved Theorem 42 for principal l-ideals. Condition (ii) below shows the concept also specializes to that of a "v-ideal" (Krull).

⁴⁶ As defined on p. 32 of the author's "Lattice Theory."

We have seen that any complemented l-ideal is the polar of its complement, and that the polar of any subset of a complete commutative l-group is a closed l-ideal; we shall now complete the circle of reasoning by showing that any closed l-ideal is complemented, yielding

THEOREM 41 (Riesz). *For any l-ideal J of a complete commutative l-group, the following assertions are equivalent: (i) J is complemented, (ii) $J = (J^*)^*$, (iii) J is closed. If (i) holds, then J^* is the complement of J .*

COMPLETION OF PROOF. If J is a closed l-ideal of any complete l-group G , then for any positive $a \in G$ we can form the J -component a_J of a , as

$$a_J = \vee_{x \in J} x \wedge a = \vee_{x \in J, x \geq 0} x \wedge a.$$

Since G is complete, and $0 \leq x \wedge a \leq a$ for all $x \geq 0$, a_J exists. Moreover since J is closed and every $x \wedge a$ is in J , a_J is in J . Hence for all positive $z \in J$, since $(z + a_J)$ is positive and in J ,

$$a_J \leq (z + a_J) \wedge a \leq \vee_{x \in J, x \geq 0} x \wedge a = a_J.$$

But now by the distributive law (1),

$$(z + a_J) \wedge a = z \wedge (a - a_J) + a_J,$$

whence, cancelling, $z \wedge (a - a_J) = 0$. Thus $a - a_J$ is in J^* . It follows that $J + J^*$ includes all positive elements $a = a_J + (a - a_J)$ of G —and hence all elements of G by Cor. 2 of Thm. 13, so that $J + J^* = G$. But evidently $J \wedge J^* = 0$, which shows that J is complemented with complement J^* , as asserted in the Theorem.

COROLLARY. *Any intersection of complemented l-ideals of a complete commutative l-group is itself complemented.*

26. Weak units and direct decompositions

Since any intersection of closed l-ideals is itself closed, it is natural to try to describe explicitly the intersection of all closed l-ideals which contain a fixed positive element a —in other words, the closed l-ideal generated by a .

We can answer this question (in complete commutative l-groups) by direct appeal to Theorem 41. Using condition (ii), we see that $(a^*)^*$ is the smallest closed l-ideal which contains a ; further, it is the largest closed l-ideal having a for weak unit. We can also describe $(a^*)^*$ in another way, using Theorem 27. Clearly any closed l-ideal which contains a will contain all x such that $x = \vee n a \wedge x = \wedge n a \vee x$; but conversely, the set of all such x is a closed l-ideal containing a . It is of course the topological closure of the principal l-ideal generated by a .

Now let A be any l-group with weak unit e . If A can be represented as the cardinal product $A_1 \cdots A_n$ of smaller l-groups, then the components of e in the different A_i are disjoint elements whose sum is e .

DEFINITION. *By a decomposition of a positive element e of an l-group A , is*

meant a set of disjoint elements e_i whose sum is e . By a component of e is meant an element e' such that⁴⁷ $e' \wedge (e - e') = 0$.

THEOREM 42. *The components of any positive element of any l-group form a Boolean algebra.*

PROOF. They are the elements x of the distributive lattice of all elements $0 \leq x \leq e$ which have complements, by definition and Theorem 13. These form a Boolean algebra, by Theorem 6.2 of the author's "Lattice Theory."

THEOREM 43. *Let A be any complete commutative l-group with weak unit e . Then the direct decompositions of A correspond one-one with the decompositions of e .*

PROOF. We have already seen that the components of e under any direct decomposition of A create a decomposition of e . But conversely, let $e = e_1 + \cdots + e_n$ be any decomposition of e , and let A_i denote the closed l-ideal generated by e_i . Since the e_i are disjoint, we will have $e_i^* \supset A_j$ if $i \neq j$, and so $A_i \perp A_j$ —or, what comes to the same thing, $A_i \wedge A_j = 0$. But the sum of the A_i contains $e_1 + \cdots + e_n$, and is a closed l-ideal; hence it contains $A = 0^* = (e^*)^*$.

This completes the proof; we note in passing that the example of the ordinal product of the additive group of the integers, with the cardinal product of the additive group of the integers with itself (in symbols, $J \circ (JJ)$), shows that the hypothesis of completeness is not redundant.

27. Residuated lattices

The concept of l-group can be generalized in two ways: one can weaken either the group or the lattice postulates. The least essential group postulate seems to be the one requiring the existence of inverses. If this is dropped, we arrive at something very close to the usual concept of a *residuated lattice*.⁴⁸

DEFINITION. *Let G be any (additively written) groupoid, or associative system with identity 0. If G is also a lattice, and satisfies*

$$(28) \quad a + \vee x_\alpha = \vee (a + x_\alpha) \quad \text{and} \quad \vee x_\alpha + b = \vee (x_\alpha + b),$$

it will be called an l-groupoid. In any l-groupoid, the left-residual $a:b$ of b by a is defined as the join of all x such that $xb \leq a$. The right-residual $a::b$ of b by a is defined as the join of all y such that $by \leq a$.

Clearly every l-group is an l-groupoid, in which $a:b$ is $a - b$ and $a::b$ is $-b + a$. Also, by (28), $(a:b) + b \leq a$ and $b + (a::b) \leq a$. The concept of l-groupoid is not self-dual; in any l-groupoid we have the monotonicity law (2), but not the dual of (28), even for finite meets.

Much of the importance of l-groupoids stems from

⁴⁷ The concepts just defined, together with Theorems 42–43, are due essentially to Freudenthal, *op. cit.*

⁴⁸ This concept, and (implicitly) that of l-groupoid, are due to M. Ward and R. P. Dilworth ("Residuated lattices," *Trans. Am. Math. Soc.* 45 (1939), pp. 335–354, and "Non-commutative residuated lattices," *ibid.* 46 (1939), pp. 426–444). The main contribution of the present section is to show that the concept applies to important systems other than ideals.

THEOREM 44 (Ward). *The ideals of any ring form an l-groupoid if inclusion is taken to mean set-inclusion and if ideal multiplication is taken as the group operation.*

We shall omit the proof, which is immediate. A special further property of ideals is $x + y \leq x \wedge y$; this is not a consequence of the postulates for an l-groupoid, and implies $x \leq 0 \wedge x$, or $0 \geq x$ for all x , as a special case. Conversely, if every $x \leq 0$, then $x + y \leq 0 + y = y$ and similarly $x + y \leq x$, whence $x + y \leq x \wedge y$, for all x, y .

DEFINITION. A residuated lattice is an l-groupoid in which $x + y \leq x \wedge y$ for all x, y ,—or equivalently, in which every element is negative.

No l-group is a residuated lattice. However, we have

THEOREM 45. *Let G be any l-group, and S any set of negative elements of G which contains 0 and is closed under $+$ and \vee . Then S is a residuated lattice.*

COROLLARY. *The set of all negative elements of any l-group or l-groupoid is a residuated lattice.*

For instance, by Theorem 45, the non-positive non-increasing real functions on any interval, the non-positive convex functions on any interval, and the non-positive subharmonic functions on any plane region, form residuated lattices.⁴⁹

THEOREM 46. *An abstract lattice L is residuated when \wedge is taken as the group operation, if and only if the dual of L is a Brouwerian logic. In this case, the residuation operation : specializes to the implication operation \rightarrow .*

PROOF: Compare the definitions given above with Theorem 8.4 of "Lattice Theory."

THEOREM 47 (J. W. Duthie⁵⁰). *The binary relations on any set form an l-groupoid, if the relative product is taken as the group operation, while the lattice operations are given their usual significance.*

We note also that the \vee -ideals of any commutative groupoid form a residuated lattice.

PROOF. The different postulates defining an l-groupoid are proved in Schroder's "Algebra der Logik," vol. III, esp. formula (29), p. 100, and formula (6), p. 79. The proof can be supplied by anyone familiar with the definitions.

The relations form a Boolean algebra under inclusion; and it has been proved by Ward-Dilworth (*op. cit.*, Thm. 7.4) that the only way to make a Boolean algebra residuated is to take lattice-meet as the group operation; hence we know in advance that relations cannot form a residuated lattice.

We note that in every l-groupoid, all left-residuals $a : a$ of elements with them-

⁴⁹ These and other function-theoretic examples of the same type were signalized in §133 of "Lattice Theory," where however the connection with residuated lattices was not remarked.

⁵⁰ Communicated to the author orally; this result was not mentioned by O. Ore in his Colloquium Lectures on relations. For the definitions of relative product, join, and meet, for binary relations, cf. A. Tarski, "Introduction to Logic," New York, 1941, pp. 90-93, or E. Schroder, "Algebra der Logik." It is interesting that the conversion operator $a \rightarrow d$ should act as an involution on the algebra of relations.

selves are idempotent; $a:a + a:a = a:a$. For by the definition of left-residual, we have

$$(a:a) + (a:a) + a \leq (a:a) + a \leq a,$$

whence $(a:a) + (a:a) \leq a:a$. Conversely, $a + 0 = a$, whence $0 \leq a:a$, and so $(a:a) = (a:a) + 0 \leq (a:a) + (a:a)$, completing the proof.—In residuated lattices, $a:a = 0$, and the result just proved is trivial.

Finally (cf. Dilworth, *op. cit.*), we can prove

$$(20'') \quad a \smile b = a \smile c = 0 \text{ implies } a \smile (b + c) = 0$$

in any l-groupoid; in fact, the proof of Theorem 15 applies as it stands!

28. Riesz' Interpolation Property⁵¹

The most basic of the lattice postulates (see §4) seem to be the reflexive law P_1 and the transitive law P_3 ; in general, a system with a reflexive and transitive relation is called a quasi-ordered set.

DEFINITION. A quasi-ordered group is a group G with a homogeneous reflexive and transitive relation. If the relation is also anti-symmetric (satisfies P_3), then G is called a partially ordered group (or semiordered group).

It is well-known ("Lattice Theory", Thm. 1.2) that in any quasi-ordered set, if $a \smile b$ is defined to mean that $a \geq b$ and $a \leq b$, we get an equivalence relation, and that we can consistently identify "equivalent" elements to get a partially ordered set. It is easily shown that in a quasi-ordered group, the $x \sim 0$ form a normal subgroup N , while the other equivalence classes form the cosets of N . This gives

THEOREM 48. The algorithm of identifying a and b whenever $a \geq b$ and $a \leq b$, yields from any quasi-ordered group a partially ordered group.

Existence postulates such as the Interpolation Properties to be discussed below and the lattice postulates L' - L'' apply to quasi-ordered groups just as well as to partially ordered groups; only uniqueness properties are lost. However, by Theorem 48, no real generality is lost if we restrict ourselves to partially ordered groups.⁵²

DEFINITION. Let m, n be any cardinal numbers. A partially ordered set will be said to have the (m, n) Interpolation Property if and only if, given x_1, \dots, x_m and y_1, \dots, y_n , such that $x_i \leq y_j$ for all i, j , we can find a z such that $x_i \leq z \leq y_j$ for all i, j .

SPECIAL CASES. The reflexive law makes the $(m, 1)$ and $(1, n)$ Interpolation Properties trivial. The $(0, 2)$ Interpolation Property is the Moore-Smith property discussed in §8; it implies the $(0, n)$ Interpolation Property for all finite n .

⁵¹ The author is greatly indebted to conversations with George Mackey and John von Neumann for material of the present section; the basic ideas go back to F. Riesz, *op. cit.*

⁵² Partially ordered groups can be bizarre enough. For instance, consider the additive group of real numbers, and let the "positive" elements be those which exceed unity; $na > 0$ need not imply $a > 0$.

The $(0, \beta)$ Interpolation Property for all β is equivalent to the existence of a universal element I , and can hold in no partially ordered group except 0 (§22, Lemma 1, Cor.). Again, the (α, β) Interpolation Property for all non-zero cardinals is equivalent to *conditional completeness*: the condition that every non-void set bounded above have a least upper bound, and dually. To see this, given any bounded set of elements x_i , form the non-void set y_j of upper bounds to the x_i ; then $x_i \leq y_j$ identically, so that z will exist with $x_i \leq y_j$ for all i, j ; by definition, $z = \vee x_i$. Similarly, the (α, β) Interpolation Property for all cardinals, zero included, is equivalent to completeness.

But, algebraically speaking, the most interesting case is the $(2, 2)$ *Riesz Interpolation Property*. By induction, this implies every (m, n) Interpolation Property with m, n finite and not zero. It is clearly weaker than the lattice property, since if $x_i \leq y_j$ for all i, j , then $x_i \leq \vee x_i \leq \wedge y_j \leq y_j$ for all i, j .

For example (F. Riesz, *op. cit.*), the polynomials, and also the rational functions with non-vanishing denominator, on any bounded closed region, form partially ordered groups which have the Riesz Interpolation Property but are not lattices.

THEOREM 49. *The following conditions on any partially ordered group G are equivalent:*

(i) *The Riesz Interpolation Property,*

(ii) *The condition of Lemma 1, §11.*

If G is commutative, they are both equivalent to

(iii) *The condition that if $a_1 + a_2 = b_1 + b_2$, and a_1, a_2, b_1, b_2 are positive, then there exist positive elements $c_{11}, c_{12}, c_{21}, c_{22}$, such that $\sum_{j=1}^2 c_{ij} = a_i$ and*

$\sum_{i=1}^2 c_{ij} = b_j$. (*Riesz Refinement Postulate*)

PROOF. First, (i) implies (ii). For if $0 \leq x, a, b \leq a + b$ then $0 \leq x, a$ and $x - b \leq x, a$ (transposing); hence if (i) holds, there exists s with $0 \leq s \leq a, s \leq x$ whence $x = s + t$ ($t \geq 0$), and $x - b \leq s$ whence $s + t \leq x \leq s + b$ and so $t \leq b$. Conversely, (ii) implies (i). By right-homogeneity, it suffices to prove that if $0, x \leq y, x + b$ then there exists s with $0, x \leq s \leq y, x + b$. But indeed, $0 \leq x + b \leq y + b$ since $x \leq y$; hence $x + b = s + t$ ($0 \leq s \leq y, 0 \leq t \leq b$), whence $x \leq x + t \leq x + b = s + t$ and $s \geq x, s \leq x + b$.

Finally, if G is commutative, then (ii) and (iii) are equivalent. For $0 \leq x \leq a + b$ ($a \geq 0, b \geq 0$) is equivalent to $a + b = x + (a + b - x)$, where all four summands are positive. To say that under these circumstances $x = s + t$ ($0 \leq s \leq a, 0 \leq t \leq b$) is equivalent to saying that $(a + b - x) = (a - s) + (b - t) = s' + t'$, where $b \geq s' \geq 0, a \geq t' \geq 0$ —whence s, s', t, t' behave as c_{ij} for (iii).

THEOREM 50. *In any partially ordered group which has the Riesz Interpolation Property, we know*

(20') $a \wedge b = 0$ and $a \wedge c = 0$ imply $a \wedge (b + c) = 0$.

PROOF. Suppose $x \leq a, b + c$. Then a and b are upper bounds to 0 and $x - c$, since $x - c \leq x \leq a$ and $b - (x - c) = (b + c) - x \geq 0$. Hence an element can be inserted between 0, $x - c$ and a, b . But since $a \wedge b = 0$, this element must be 0. Hence $x - c \leq 0, x \leq c$ as well as $x \leq a$, and $x = 0$.

By duality (20'') holds also; for the further study of commutative partially ordered groups with the Riesz Interpolation Property, with especial emphasis on the linear functionals on such groups, see F. Riesz, *op. cit.*

29. Unsolved problems, general case

We shall conclude this paper with a list of problems of varying degrees of interest and difficulty. For the purpose of classification, these will be divided into those which involve general l-groups, and those which relate primarily to commutative l-groups.

PROBLEM 1. Show that $na > nb$ for one positive n implies $a > b$.

SUGGESTIONS. This is easy in the simply ordered case, or if a and b are permutable (see §9, Lemma 3).

PROBLEM 2. Show that if $a > 0$ and $b > 0$, then

$$-a - b + a + b \ll a + b.$$

In words, the commutator of a and b is incomparably smaller than $a + b$.

SUGGESTION. If the commutator is in the center, then

$$n(a + b) \geq -na - nb + n(a + b) = \binom{n}{2}(-a - b + a + b).$$

Hence if the conjecture of Problem 2 can be proved, we have $(a + b) \geq 1/2n(-a - b + a + b)$ for every even integer n , giving the desired result. This method, with the aid of finite induction, might be successfully applied at least to hypercentral l-groups.

PROBLEM 3. Prove that every Archimedean l-group is commutative.

This result would be a corollary of the result conjectured in Problem 2. Using Theorem 38, we would infer as a second corollary that every *complete l-group was commutative*. Hence to disprove the conjectures of Problems 3-4, it would be enough to find a complete non-commutative l-group, or an Archimedean non-commutative l-group.

PROBLEM 4. Prove that a complete l-group either satisfies the chain condition or has at least the cardinal number of the continuum.

PROBLEM 5. Find an l-group without proper l-ideals which is non-commutative.

SUGGESTIONS. By Theorem 30, it would suffice to find a simple l-group which was non-Archimedean or not simply ordered. By Theorem 25, this is also necessary, so that if the author's conjecture is correct, either a non-Archimedean or a non-simply ordered simple l-group must exist. The author conjectures that the former is certainly the case.

PROBLEM 6. Find all l-group orderings (homogeneous lattice orderings) of

the free group with two generators. Is the commutator-subgroup necessarily an \mathfrak{l} -ideal?

SUGGESTION. See Problem 2.

PROBLEM 7. Find a necessary and sufficient condition that an abstract group be isomorphic with the additive group of an \mathfrak{l} -group.

SUGGESTION. By Theorem 16, it is necessary that every element be of infinite order; by Theorem 24, this is sufficient in the commutative case; the author conjectures that it is also sufficient in the hypercentral case.

PROBLEM 8. Suppose that in an \mathfrak{l} -groupoid $0:(0:x) = x$ and $0:: (0::x) = x$ for all x . What can be inferred?

SUGGESTIONS. The correspondence $x \rightarrow 0:x$ will then be a lattice involution, so that the dual of (28) also holds. What about the commutative case? Will $0:(x + y) = 0:x + 0:y$?

30. Unsolved problems, commutative case

Any commutative group without elements of finite order whose cardinal number is at most that of the continuum, is isomorphic with an additive subgroup of the ordered group of real numbers under addition—proof by rational bases,—and so is isomorphic with an Archimedean \mathfrak{l} -group.

PROBLEM 9. Is every commutative group without elements of finite order isomorphic with the additive group of an *Archimedean* \mathfrak{l} -group?

PROBLEM 10. Find a necessary and sufficient condition that a commutative partially order group be group- and order-isomorphic with an additive subgroup of a cardinal product of simply ordered *Archimedean* \mathfrak{l} -groups—or equivalently, by real functions.⁵³

PROBLEM 11. Given \mathfrak{l} -groups B and C , reduce the problem of finding all \mathfrak{l} -groups A having an \mathfrak{l} -ideal J isomorphic with C and \mathfrak{l} -quotient-group A/J isomorphic with B to a problem in pure group extension, in the commutative case.

This problem was implicitly solved in special cases in §19; the special cases B simple and C simple might well be attacked first.

PROBLEM 12. Find all Lie \mathfrak{l} -algebras: Lie algebras over the real field which are vector lattices relative to a set of positive elements which is invariant under all inner automorphisms.

SUGGESTIONS. Use the known classification of vector lattices with finite basis (Cors. 1–2 of Theorem 35). The author conjectures that a Lie algebra can be made into a Lie \mathfrak{l} -algebra only if it is *solvable* (or “integrable”).

PROBLEM 13. Find all Lie \mathfrak{l} -groups in the large.

The problem in the small is contained in Problem 12; the fact that all elements have infinite order should simplify it.

PROBLEM 14. Construct a theory of \mathfrak{l} -rings.

⁵³ The work of von Neumann (unpublished), Stone (cf. footnote 35) et al. shows that any such Archimedean group is isomorphic with a homomorphic image of such a cardinal product.

The only postulates known (Stone, *op. cit.*) cover only a very special case: subrings of cardinal unions of simply ordered l-rings, corresponding to rings of functions. Cf. also A. A. Albert, "On ordered algebras", Bull. Am. Math. Soc. 46 (1940), pp. 521-522.

PROBLEM 15. Find a more direct substitute for condition (8) in Theorem 9: i.e., a simple condition or set of simple conditions on the operation $a \rightarrow a^+$ necessary and sufficient to make the operation $(a - b)^+ + b$ associative.

HARVARD UNIVERSITY

OPERATOR METHODS IN CLASSICAL MECHANICS, II

By PAUL R. HALMOS AND JOHN VON NEUMANN

(Received December 23, 1941)

Introduction

The purpose of this paper is two-fold: to map all measure spaces for which this is possible on the unit interval, and to apply such mapping theorems to the study of ergodic measure preserving transformations with a pure point spectrum.

"Mappings" between two measure spaces may be interpreted in two ways, as set mappings and as point mappings, and accordingly we give below two sets of necessary and sufficient conditions for the existence of a mapping from a given space to the interval. The first of these, the set mapping or algebraic isomorphism theorem, seems to be known, and although it has never been explicitly stated in the literature there are many proofs of special cases of it on record. We give an explicit proof of it and use a construction of the proof in proving the second, point mapping or geometric isomorphism, theorem. This second theorem depends on the new concept of normal measure space: a seemingly artificial concept which is, however, useful for two reasons. First, it is purely measure theoretic (and not topological), in character, and hence is applicable to the measure spaces usually discussed in probability theory; second it is hereditary under all the usual operations on measure spaces (such as the formation of direct products, decomposition into direct sums, etc.).

Using the concepts and results of the mapping theorems just described, and of the Pontrjagin duality theorem concerning compact and discrete abelian groups, we are able to show that every ergodic measure preserving transformation with a pure point spectrum is isomorphic to a rotation on a compact abelian group. This is a "normal form" theorem for a certain class of measure preserving transformations and can be used to answer many questions, such as the existence of square roots, commutative transformations, etc., concerning such transformations.

Although this paper is a continuation of an earlier work of one of us¹ it is to a large extent independent of this earlier work. The proofs of the main theorems mentioned above are logically complete here; only in some of the applications, as for example in discussing the relation between point mappings and set mappings, do we make use of the results of (I).

1. General measure spaces; the algebraic isomorphism theorem

Let X be any set, and \mathfrak{C} any Borel field of subsets of X ; let m be a non negative, countably additive, finite measure defined on \mathfrak{C} . The system $\{X, \mathfrak{C}, m\}$, which we shall usually denote by X , or, if necessary to indicate its

¹ See John von Neuman, *Zur Operatorenmethode in der klassischen Mechanik*, *Annals of Mathematics*, vol. 33, (1932), pp. 587-642. In the sequel we shall refer to this paper as (I).

dependence on \mathcal{X} and m , by $X(\mathcal{X}, m)$ is called a *measure space*. Sets $E \in \mathcal{X}$ are called *measurable*; we shall use also the usual terminology of the Lebesgue theory in describing functions as measurable, integrable, etc. A measure space is *complete* if every subset of a measurable set of measure zero is itself measurable (and has, of course, measure zero). Since it is always possible to extend the definition of m to a Borel field $\mathcal{X}' \supset \mathcal{X}$ so that $X(\mathcal{X}', m)$ is complete, we shall lose no generality, and gain somewhat in simplicity, by assuming completeness.

In any measure space X we shall write $\mathfrak{B} = \mathfrak{B}(\mathcal{X})$ for the Boolean algebra of measurable sets modulo sets of measure zero. We shall make use of the notations of set theory, (\subset , $+$, etc.) in \mathfrak{B} , and of the fact that we may consider m as defined on \mathfrak{B} .

We discuss now the concept of separability in measure spaces. A Borel field \mathcal{X} (or a measure space $X(\mathcal{X}, m)$) is *strictly separable* if it contains a countable collection of sets such that the smallest Borel field containing all of them, (the Borel field *spanned* by them), is \mathcal{X} itself. Two sub Borel fields, \mathcal{A} and \mathcal{B} , of the Borel field \mathcal{X} of measurable sets in a measure space $X(\mathcal{X}, m)$ are *equivalent* if to every set E in either one of them there corresponds a set F in the other such that the symmetric difference $(E - F) + (F - E)$ has measure zero. A measure space is *separable* if there exists a strictly separable Borel field \mathcal{A} contained in and equivalent to \mathcal{X} .² A concept, which lies logically between separability and strict separability, more useful than either of these, is *proper separability*. A measure space $X(\mathcal{X}, m)$ is *properly separable* if there exists a strictly separable Borel field $\mathcal{A} \subset \mathcal{X}$, such that to every $E \in \mathcal{X}$ there corresponds an $F \in \mathcal{A}$ with $E \subset F$ and $m(F - E) = 0$.³ We observe that this definition is self dual: by applying the condition to $X - E$ we readily obtain a set $F \in \mathcal{A}$ with $F \subset E$ and $m(E - F) = 0$. We shall make use of the fact that if X is separable (or properly separable) and \mathcal{A} is the strictly separable Borel field described in the definitions above then $\mathfrak{B}(\mathcal{X}) = \mathfrak{B}(\mathcal{A})$. In the case of (properly) separable measure spaces it will be necessary to indicate in the notation the strictly separable Borel field used; we shall write $X = X(\mathcal{X}, \mathcal{A}, m)$. We shall call sets of \mathcal{A} *Borel sets*, and functions measurable (\mathcal{A}) *Baire functions*. (A real valued function $f(x)$ is measurable (\mathcal{A}) if the inverse image under f of every real Borel set S , i.e. the set $\{x \mid f(x) \in S\}$, belongs to \mathcal{A} .)

² This is not the usual form in which this definition is given. Cf., for example, J. L. Dobb, *One-parameter families of transformation*, Duke Mathematical Journal, vol. 4, (1938), p. 753. That our definition is, however, equivalent to the usual one is proved by Paul R. Halmos, *The decomposition of measures*, Duke Mathematical Journal, vol. 8, (1941), p. 387. We observe X is separable if and only if the Boolean algebra $\mathfrak{B}(\mathcal{X})$ has a countable number of generators.

³ The concept of proper separability, first introduced by W. Ambrose and S. Kakutani, *Structure and continuity of measurable flows*, Duke Mathematical Journal, vol. 9, (1942), pp. 25-42, is fundamental in measure theory. Although it is possible to give examples of separable but not properly separable measure spaces, these examples are all of a more or less pathological kind. One such example is the unit interval, with the Borel field of all sets of Lebesgue measure zero and their complements in the role of \mathcal{X} .

A measurable set E in the measure space $X(\mathfrak{C}, m)$ is *indecomposable* if it contains no proper measurable subsets other than the empty set; an element $E \in \mathfrak{B}(\mathfrak{C})$ is an *atom* if it contains no proper subelements, other than 0, in $\mathfrak{B}(\mathfrak{C})$. A measure space is *non atomic* if $\mathfrak{B}(\mathfrak{C})$ has no atoms: in other words if every measurable set of positive measure contains measurable subsets of smaller positive measure. From the point of view of a study of the structure of measure spaces indecomposable sets and atoms are uninteresting: we shall generally assume that the former consist of exactly one point and the latter are absent. More specifically our assumption will be described in the following terms.

A countable sequence, A_1, A_2, \dots , of subsets of X is a *separating sequence* if to every pair of points, $x \neq y$, we may find an integer n with $x \in A_n$, $y \in X - A_n$. If there exists in X a separating sequence of measurable sets, an indecomposable set contains exactly one point. We shall now show that the assumption of the existence of a separating sequence of measurable sets has a similar effect on atoms. Let E be a set of positive measure which contains no measurable subsets of smaller positive measure. It follows that for each n one of the two sets, EA_n , and $E(X - A_n)$ has measure zero and the other one has measure $m(E)$. By a slight change of notation we may assume $m(EA_n) = m(E)$ for $n = 1, 2, \dots$. If we write $\prod_{n=1}^{\infty} A_n = A$, then we have $m(EA) = m(E)$; since, however, A can contain at most one point, this implies that for some point $x \in E$ we have $m(E - x) = 0$. In other words the existence of a measurable separating sequence implies that the weight of an atom is concentrated at one point; if, for example, we assume that the measure of a point is always zero, we may infer that the space is non atomic. Since in a measure space, which has by definition finite measure, there can be at most a countable set of points of positive measure, and since their measure theoretic structure is clear, we shall generally assume non-atomicity explicitly.

If $X_1(\mathfrak{C}_1, m_1)$ and $X_2(\mathfrak{C}_2, m_2)$ are measure spaces, a *set isomorphism* between X_1 and X_2 is a measure preserving isomorphism between the Boolean algebras $\mathfrak{B}(\mathfrak{C}_1)$ and $\mathfrak{B}(\mathfrak{C}_2)$. More specifically a set isomorphism is a one to one mapping T from $\mathfrak{B}(\mathfrak{C}_1)$ on $\mathfrak{B}(\mathfrak{C}_2)$ which is such that

$$\begin{aligned} T(X_1 - E) &= X_2 - TE, \\ T(\sum_{n=1}^{\infty} E_n) &= \sum_{n=1}^{\infty} TE_n, \\ m_1(E) &= m_2(TE). \end{aligned}$$

If such a mapping T exists, X_1 and X_2 are *set isomorphic*.

After one more comment on notation we shall be ready to state and prove our first result. Since the unit interval plays a fundamental role in our investigations and is used as a yardstick with which to compare other measure spaces, we find it convenient to introduce a special notation for it. We shall denote the unit interval by \tilde{X} , the collection of Lebesgue and Borel measurable sets by

\mathfrak{X} and $\tilde{\mathfrak{Q}}$ respectively, and Lebesgue measure by \tilde{m} . In our terminology $\tilde{X} = \tilde{X}(\mathfrak{X}, \tilde{\mathfrak{Q}}, \tilde{m})$ is a properly separable measure space.⁴

THEOREM 1. *A necessary and sufficient condition that a measure space of total measure one be set isomorphic to the unit interval is that it be separable and non-atomic.*

PROOF. Since the unit interval is separable and non-atomic and since these properties are evidently invariant under set isomorphisms, the necessity of our conditions is clear. To prove their sufficiency, let $X(\mathfrak{X}, \mathfrak{Q}, m)$ be the given measure space, $m(X) = 1$, and let A_1, A_2, \dots be a countable sequence of Borel sets which span \mathfrak{Q} . We may assume (by adding a superfluous set to the $\{A_n\}$ if necessary) that $\sum_{n=1}^{\infty} A_n = X$. Then we may make correspond to every rational number r , $0 \leq r \leq 1$, a set B_r such that

(i) $\{A_n\}$ and $\{B_r\}$ span the same field;

(ii) $r < s$ implies $B_r \subset B_s$;

(iii) $\prod_{r>a} B_r = B_a$;

(iv) $\prod_r B_r = 0$; $\sum_r B_r = X$.⁵

We now define, for every real number a , $0 \leq a \leq 1$, a set B_a by $B_a = \prod_{r>a} B_r$. It is clear that this definition of B_a is consistent with its previous definition in case a is rational, and that the family of sets $\{B_a\}$ satisfies the conditions (ii), (iii), (iv), (where in (iii) and (iv) we extend the products and sums over an arbitrary countable set of real numbers r for which $\inf r = s$ in (iii), $\inf r = 0$ and $\sup r = 1$, respectively, in (iv)). Moreover, condition (i) implies that $B_a \in \mathfrak{Q}$ for all a and that the Borel field spanned by the B_a is \mathfrak{Q} itself.

Given now the family B_a we may find a (uniquely determined) function $f(x)$, defined for $x \in X$, $0 \leq f(x) \leq 1$, for which $\{x \mid f(x) \leq a\} = B_a$; we may, for example, define

$$(1) \quad f(x) = \inf \{a \mid x \in B_a\}.$$

The class of all sets of the form

$$(2) \quad f^{-1}(\tilde{E}) = \{x \mid f(x) \in \tilde{E}\},$$

where \tilde{E} is an arbitrary Borel set in the unit interval, is a Borel field contained in \mathfrak{Q} ; since it contains all B_a , and therefore all A_n , it coincides with \mathfrak{Q} .

Let $F(a) = m\{x \mid f(x) \leq a\} = m(B_a)$ be the distribution function of $f(x)$: $F(a)$ is monotone non-decreasing from 0 to 1 as a ranges between 0 and 1, and is continuous from the right. (This much is always true, of an arbitrary distribution function.) In our special case we assert that $F(a)$ is continuous. For

⁴ In the sequel we shall sometimes use the notation $\tilde{X}(\mathfrak{X}, \tilde{\mathfrak{Q}}, \tilde{m})$ for the perimeter of the unit circle in the complex plane: it is clear that this space has the same measure theoretic structure as the unit interval. We shall always make it clear whether the symbol \tilde{X} has its real or its complex meaning.

⁵ Cf. (I), p. 602; see also J. L. Doob, *Stochastic processes with an integral valued parameter*, Transactions of the American Mathematical Society, vol. 44, (1938), p. 91.

if $a = a_0$ is a discontinuity of $F(a)$, then $\{x \mid f(x) = a_0\}$ is a set of positive measure which therefore, (non-atomicity), has Borel subsets of smaller positive measure. Such a subset cannot be put in the form $f^{-1}(\bar{E})$, contrary to what we have already proved.

For any \bar{x} , $0 \leq \bar{x} \leq 1$, we define $\bar{f}(\bar{x}) = \inf \{a \mid F(a) \geq \bar{x}\}$. It is well known (and easily verified) that $\bar{f}(\bar{x})$ is a strictly monotone increasing (not necessarily continuous) function of \bar{x} , which increases from 0 to 1 as \bar{x} does, and which is continuous on the left. Moreover the distribution function of $\bar{f}(\bar{x})$ is again $F(a)$.

For any Borel set $\bar{E} \subset \bar{X}$, consider the set $\bar{f}^{-1}(\bar{E})$: we assert that the collection of all sets of this form, (which clearly forms a Borel field), coincides with $\bar{\mathcal{A}}$. This is true since the increasing character of $\bar{f}(\bar{x})$ implies that every interval $(0, \bar{x})$ has the form $\bar{f}^{-1}(\bar{E})$, where \bar{E} can even be chosen as an interval.

Suppose that it ever happens that $\bar{f}^{-1}(\bar{E}_1) = \bar{f}^{-1}(\bar{E}_2)$. (We shall now make use of the fact that for an arbitrary Baire function $g(x)$, $0 \leq g(x) \leq 1$, the correspondence $\bar{E} \rightarrow g^{-1}(\bar{E}) = \{x \mid g(x) \in \bar{E}\}$, is a homomorphism of $\bar{\mathcal{A}}$ into \mathcal{A} , i.e. that $g^{-1}(\bar{X} - \bar{E}) = X - g^{-1}(\bar{E})$, and $g^{-1}(\bar{E}_1 + \bar{E}_2 + \dots) = g^{-1}(\bar{E}_1) + g^{-1}(\bar{E}_2) + \dots$). If we write $\bar{E}' = (\bar{E}_1 - \bar{E}_2) + (\bar{E}_2 - \bar{E}_1)$ for the symmetric difference between \bar{E}_1 and \bar{E}_2 , then it follows from the equality of the distributions of $f(x)$ and $\bar{f}(\bar{x})$, that $m\{f^{-1}(\bar{E}')\} = \bar{m}\{\bar{f}^{-1}(\bar{E}')\} = 0$. Conversely, of course, $\bar{f}^{-1}(\bar{E}_1) = \bar{f}^{-1}(\bar{E}_2)$ implies the same result.

Consequently the correspondence $\bar{f}^{-1}(\bar{E}) \rightleftharpoons \bar{f}^{-1}(\bar{E})$ is one to one, not necessarily between \mathcal{A} and $\bar{\mathcal{A}}$, but certainly between $\mathfrak{B} = \mathfrak{B}(\mathcal{X}) = \mathfrak{B}(\mathcal{A})$ and $\bar{\mathfrak{B}} = \mathfrak{B}(\bar{\mathcal{X}}) = \mathfrak{B}(\bar{\mathcal{A}})$. It is clear that this correspondence preserves measure, and the homomorphic nature of the mappings $\bar{E} \rightarrow f^{-1}(\bar{E})$ and $\bar{E} \rightarrow \bar{f}^{-1}(\bar{E})$ shows that it is also an algebraic isomorphism.

This concludes the proof of Theorem 1.

2. Normal spaces; the geometric isomorphism theorem

If $X_1(\mathcal{X}_1, m_1)$ and $X_2(\mathcal{X}_2, m_2)$ are measure spaces, a *point isomorphism* between X_1 and X_2 is a one to one mapping from almost all of X_1 on almost all of X_2 such that $E_1 \in \mathcal{X}_1$ if and only if $E_2 = TE_1 \in \mathcal{X}_2$, and then $m_1(E_1) = m_2(E_2)$. If such a mapping T exists, X_1 and X_2 are *point isomorphic*. Our problem in this section is to find necessary and sufficient conditions in order that a measure space be point isomorphic to the unit interval. The fundamental concept in this connection is that of a normal space.

DEFINITION 1. A measure space is proper if it is complete, properly separable, and non-atomic, and if it contains a separating sequence of Borel sets.

DEFINITION 2. A proper measure space is normal if to each real valued univalent Baire function $f(x)$ there corresponds a set X_0 of measure zero such that the range, $f(X - X_0)$, is a Borel set.

The following lemmas concerning proper and normal spaces will be useful in the sequel.

LEMMA 1. On every proper measure space $X(\mathcal{X}, \mathcal{A}, m)$ there exist real valued bounded univalent Baire functions.

PROOF. Since X is certainly separable and non-atomic the construction of the proof of Theorem 1 applies. We assert that the real valued bounded Baire function $f(x)$ defined by (1) is univalent. For if the set $\{x \mid f(x) = a\}$ contained more than one point, then the intersection of this set with a Borel set separating two of its points could not be expressed in the form $\{x \mid f(x) \in \tilde{E}\}$. Since, however, the proof of Theorem 1 establishes that every Borel set has this form, $f(x)$ must be univalent.

LEMMA 2. *If $X(\mathcal{C}, \mathcal{A}, m)$ is a proper measure space with the property that the condition of Definition 2 is satisfied by every bounded function then X is normal.*

PROOF. Let $f(x)$ be any univalent Baire function, and let $G(y)$ be any continuous function which maps the infinite interval, $-\infty < y < +\infty$, in a one to one way on a finite interval. Then $g(x) = G(f(x))$ is a Baire function which is univalent and bounded, hence, by hypothesis, there is a set X_0 of measure zero such that $g(X - X_0)$ is a Borel set. The image of this Borel set under the one to one continuous mapping $G^{-1}(y)$ is the range $f(X - X_0)$ which is therefore also a Borel set.

LEMMA 3. *If $X(\mathcal{C}, \mathcal{A}, m)$ is a normal space, $B \subset X$ is a Borel set, and $f(x)$ is a real valued univalent Baire function, then there is a set $B_0 \subset B$ of measure zero such that $f(B - B_0)$ is a Borel set. B_0 can even be chosen in the form BX'_0 , where X'_0 is a Borel set of measure zero, depending on f but not on B .*

PROOF. We shall carry out the proof in three steps, first establishing the existence of a suitable B_0 corresponding to a fixed B , then showing that B_0 may even be chosen as a Borel set, and, finally, proving on the basis of our separability hypotheses, that we may choose B_0 in the form described in the statement of the lemma.

(i) We observe that the first statement asserts, essentially, that a Borel set in a normal space is itself a normal space. Accordingly, using Lemma 2, we may assume that $f(x)$ is bounded. Let $f'(x)$ be a bounded univalent Baire function on X , (Lemma 1); by appropriate linear transformations of $f(x)$ and of $f'(x)$ we can secure

$$0 \leq f(x) \leq 1 < f'(x)$$

throughout X . Then the function $f^*(x)$, defined to be equal to $f(x)$ on B and to $f'(x)$ on $B' = X - B$ is a univalent Baire function on X , hence for a suitable set X_0 of measure zero, $f^*(X - X_0)$ is a Borel set. The intersection of this Borel set with the closed interval $(0, 1)$ is also a Borel set: this intersection is, however, precisely $f(B - B_0)$, where $B_0 = BX_0$.

(ii) Let B_1 be a Borel set of measure zero, $B_1 \supset B_0$. Applying the result of (i) to $X - B_1$ we may find a set $B_2 \supset B_1$ of measure zero such that $f(X - B_2)$ is a Borel set. We proceed similarly by induction, choosing $B_3 \supset B_2$ to be a Borel set of measure zero, choosing $B_4 \supset B_3$ so that $f(X - B_4)$ is a Borel set, and so on. We have $B_0 \subset B_1 \subset B_2 \subset B_3 \subset \dots$; all B_n are of measure zero; or n odd B_n is a Borel set; for n even $f(X - B_n)$ is a Borel set. We write $B_0^* = \sum_{n=0}^{\infty} B_n$. Then B_0^* has measure zero, and, because of the monotone

character of the sequence $\{B_n\}$, $B_0^* = \sum_{n=0}^{\infty} B_{2n+1}$, so that B_0^* is a Borel set. Similarly $X - B_0^* = X - \sum_{n=0}^{\infty} B_{2n} = \prod_{n=0}^{\infty} (X - B_{2n})$, so that $f(X - B_0^*)$ is a Borel set, and also $B(X - B_0^*) = (B - B_0)(X - B_0^*)$, so that $f(B(X - B_0^*)) = f(B - B_0)f(X - B_0^*)$ is a Borel set. We may accordingly change notation and denote by B_0 the intersection of B and B_0^* : this new B_0 is a Borel set of measure zero with the property that $f(B - B_0)$ is a Borel set.

(iii) Let A_1, A_2, \dots be a sequence which spans \mathcal{G} , and apply the result of (ii) to find, for each n , a Borel set $A_n^0 \subset A_n$, of measure zero, such that $f(A_n - A_n^0)$ is a Borel set. We write $A^0 = \sum_{n=1}^{\infty} A_n^0$, and we apply (ii) once more, this time to $X - A^0$, to find a Borel set $X'_0 \supset A^0$, of measure zero, such that $f(X - X'_0)$ is a Borel set. Let us write $A'_n = A_n - A_n^0$, and let \mathcal{G}' be the Borel field ($\subset \mathcal{G}$) spanned by the A'_n . Then we have $(X - X'_0)A_n = (X - X'_0)A'_n$ for all n , and we see, moreover, that to every Borel set B , (i.e. to every set $B \in \mathcal{G}$), there corresponds a set $B' \in \mathcal{G}'$ such that $(X - X'_0)B = (X - X'_0)B'$. Since $f(A'_n)$ is a Borel set, and since the collection of sets A for which $f(A)$ is a Borel set is clearly a Borel field, (because f is univalent), it follows that for every $B' \in \mathcal{G}'$, $f(B')$ is a Borel set. Consequently for every $B \in \mathcal{G}$

$$f(B - BX'_0) = f(B(X - X'_0)) = f(B'(X - X'_0)) = f(B')f(X - X'_0),$$

so that $f(B - BX'_0)$ is a Borel set, and the proof of the lemma is complete.

LEMMA 4. If $X(\mathcal{X}, \mathcal{G}, m)$ is a proper measure space, and if for a single real valued univalent Baire function $g(x)$ we can find a set X_0 of measure zero such that $g(B - BX_0)$ is a Borel set whenever B is, then X is normal and, moreover, this same set X_0 will satisfy the condition of definition 2 for any real valued univalent Baire function $f(x)$.

PROOF. We write $Y = g(X - X_0)$; for every $y_0 \in Y$, $y_0 = g(x_0)$, we define $F(y_0) = f(x_0)$. $F(y)$ is then a real valued univalent function of the real variable $y \in Y$. Since

$$(3) \quad \{y \mid F(y) < a\} = g[\{x \mid f(x) < a\}(X - X_0)],$$

and since the right member is a Borel set by hypothesis, $F(y)$ is a Baire function. Since $f(x) = F(g(x))$, we have $f(X - X_0) = F(Y)$, and therefore $f(X - X_0)$ is a Borel set.⁶

An important class of measure spaces is the class of m -spaces. An m -space is a complete measure space $X(\mathcal{X}, m)$ on which a metric is defined so that, topologically, it is a complete separable space, and which satisfies the following two conditions:

(i) the measure of an open set is positive;

(ii) for every measurable set E , $m(E) = \inf \{m(O) \mid E \subset O, O \text{ open}\}$. With the Borel field \mathcal{G} of Borel sets (in the usual topological sense of the word) $X = X(\mathcal{X}, \mathcal{G}, m)$ becomes a proper measure space; it is a known result of topology

⁶ See F. Hausdorff, *Mengenlehre*, Berlin, 1935, p. 266.

that it is even normal in our sense of the word, and that the exceptional set X_0 of measure zero may even be chosen as the empty set.⁷

We shall use m -spaces later; at present we mention them only as examples of normal spaces. The following theorem, the main theorem of the present section, applies to m -spaces, (since they are normal), and shows that, measure theoretically, they are isomorphic to the unit interval.

THEOREM 2. *A necessary and sufficient condition that a measure space of total measure one be point isomorphic to the unit interval is that it be normal.*

PROOF. The necessity of our condition is obvious: the unit interval is normal and normality is invariant under point isomorphism. Before giving a proof of sufficiency we remark on the hypotheses. Since the various conditions in the definition of a *proper* space are logically independent, they are obviously indispensable for a sufficiency proof. It is possible that the condition of *normality* could be replaced by a weaker one, but examples seem to indicate that it is the best way of expressing that the space is "measurable in itself."

For the proof of sufficiency we use the notations of the proof of Theorem 1; in particular we use the functions $f(x)$ and $\tilde{f}(\tilde{x})$ that we defined there.

We denote by D and \tilde{D} the ranges of $f(x)$ and $\tilde{f}(\tilde{x})$ respectively. By omitting from X a set of measure zero we may, by normality, assume that D is a Borel set; \tilde{D} is also a Borel set. (We observe that the omission of a set of measure zero does not change the distribution of f and hence does not change \tilde{f} at all). Form the set $R = (D - \tilde{D}) + (\tilde{D} - D)$. Since $f^{-1}(\tilde{D} - D) = f^{-1}(\tilde{D}) - f^{-1}(D)$ lies entirely in the complement of $f^{-1}(D)$, and since this complement is empty, $f^{-1}(\tilde{D} - D)$ is empty. Since $\tilde{f}^{-1}(\tilde{D} - D)$ has the same measure as $f^{-1}(\tilde{D} - D)$, this proves that the measure of $\tilde{f}^{-1}(\tilde{D} - D)$ is zero. Similarly we can prove that the measure of both $f^{-1}(D - \tilde{D})$ and $\tilde{f}^{-1}(D - \tilde{D})$ is zero, (and, in fact, the latter is empty). Hence if we omit from both X and \tilde{X} a Borel set, namely $f^{-1}(R)$ and $\tilde{f}^{-1}(\tilde{R})$ respectively, of measure zero, on the remainder f and \tilde{f} are univalent Baire functions with identical (Borel measurable) ranges.

If to every $x \in X$ (after the omission, as described, of a set of measure zero), we make correspond the point $\tilde{f}^{-1}(f(x)) \in \tilde{X}$, the correspondence is one to one. Moreover if B is any Borel set in X , and $B' = f(B)$, then B' is a Borel set and $f^{-1}(B') = B$. Consequently, considered as an element of the Boolean algebra $\mathfrak{B}(\mathfrak{X})$, the correspondent, under the set mapping described in the proof of theorem 1, of B is $\tilde{f}^{-1}(B') = \tilde{B} = \tilde{f}^{-1}(f(B))$, so that the point mapping just described induces precisely the same set isomorphism between \mathfrak{B} and \mathfrak{B} . It follows that this point correspondence is measure preserving. This concludes the proof of Theorem 2.

3. The relation between set transformations and point transformations

If T is a measure preserving transformation (i.e. a point isomorphism) of a measure space $X(\mathfrak{X}, m)$ on itself, then T induces a set mapping (of $\mathfrak{B} = \mathfrak{B}(\mathfrak{X})$

⁷ See Hausdorff, op. cit., p. 269.

on itself) by making correspond to every set $E \in \mathfrak{C}$ the set $TE \in \mathfrak{C}$. It is known that in an m -space the converse is true: every set isomorphism is induced in this way by a point isomorphism.⁸ Motivated by this we give the following definition.

DEFINITION 3. A measure space $X(\mathfrak{C}, m)$ has sufficiently many measure preserving transformations if every set isomorphism of \mathfrak{B} on itself is induced by a point isomorphism of X on itself.

It follows from Theorem 2 that every normal space has sufficiently many measure preserving transformations. In between the two concepts (normal spaces and spaces with sufficiently many measure preserving transformations) there is, however, room for a pathological occurrence which we shall describe in this section. We begin by proving some auxiliary results.

LEMMA 5. If two point mappings, on a measure space X which contains a separating sequence E_1, E_2, \dots of measurable sets, induce the same set mapping on \mathfrak{B} then they differ on at most a set of measure zero.

PROOF. It is sufficient to consider the case where one of the transformations is the identity. If then TE_n and E_n differ only on a set of measure zero, for $n = 1, 2, \dots$, it follows that all $T^k E_n$ differ from each other only on sets of measure zero. Hence the invariant set

$$F_n = \sum_{k=-\infty}^{\infty} T^k E_n - \prod_{k=-\infty}^{\infty} T^k E_n$$

has measure zero. We form the invariant set X' by omitting from X the set $\sum_{n=1}^{\infty} F_n$ of measure zero. If now $x \neq Tx$, then some E_n contains one but not both of x and Tx , and therefore x is contained in one but not both of E_n and $T^{-1}E_n$. Consequently $x \in F_n$, so that $x \notin X'$.

LEMMA 6. Let $X(\mathfrak{C}, m)$ be a measure space and let $X' \subset X$ be any (not necessarily measurable) subset of X . Let \mathfrak{C}' be the collection of all sets of the form $E' = X'E$, with $E \in \mathfrak{C}$; for every $E' \in \mathfrak{C}'$, $E' = X'E$, define $m'(E') = m(E)$. With these definitions m' is uniquely determined (so that $X'(\mathfrak{C}', m')$ is a measure space) if and only if the outer measure of X' in X is equal to the measure of X .⁹

LEMMA 7. If $\{\phi_n(x)\}$, $n = 1, 2, \dots$, is a complete orthonormal set of functions in $L_2(X)$, where $X(\mathfrak{C}, m)$ is a measure space which contains a separating sequence, E_1, E_2, \dots , of measurable sets, then there is a set $N \in \mathfrak{C}$ of measure zero such that $x, y \notin N$ and $\phi_n(x) = \phi_n(y)$ for $n = 1, 2, \dots$, implies $x = y$.

⁸ See John von Neumann, *Einige Sätze über messbare Abbildungen*, Annals of Mathematics, vol. 33, (1932), p. 582. In definition 5, p. 576, all descriptive properties of the transformation (such for example as $M_1 + M_2 \rightarrow M'_1 + M'_2$) should be modified by the phrase "neglecting sets of measure zero."

⁹ The outer measure of E_0 , $m^*(E_0)$, is defined by $m^*(E_0) = \inf \{m(E) \mid E_0 \subset E \in \mathfrak{C}\}$. Similarly we may define the inner measure, $m_*(E_0) = \sup \{m(E) \mid E_0 \supset E \in \mathfrak{C}\}$. If X is complete then E_0 is measurable (i.e. $E_0 \in \mathfrak{C}$) if and only if $m_*(E_0) = m^*(E_0) = m^*(E_0) = m(E_0)$. In case X is properly separable it is sufficient to take the supremum and infimum over Borel sets E . For the proof of Lemma 6, see J. L. Doob, *Stochastic processes depending on a continuous parameter*, Transactions of the American Mathematical Society, vol. 42, (1937), pp. 109-110.

PROOF. Let $\psi_m(x)$ be the characteristic function of E_m ; we have

$$(4) \quad \psi_m(x) = \sum_{n=1}^{\infty} a_{nm} \phi_n(x),$$

in the sense of convergence in the mean (or order two). Consequently, for each m , a subsequence of the partial sums of the series in (4) converges to $\psi_m(x)$ almost everywhere; for each m we choose a fixed subsequence with this property and we let N be the union of all the sets of measure zero at which these subsequences do not converge to $\psi_m(x)$. If $x, y \notin N$ and $\phi_n(x) = \phi_n(y)$ for all n , then it follows that $\psi_m(x) = \psi_m(y)$ for all m , whence (using the fact that E_1, E_2, \dots is a separating sequence) $x = y$.

LEMMA 8. Let $\tilde{X}(\tilde{\mathcal{X}}, \tilde{\mathcal{Q}}, \tilde{m})$ be the perimeter of the unit circle in the complex plane, and let $\mu(\tilde{E})$ be any measure (i.e. a countably additive, non-negative set function with $\mu(\tilde{X}) = 1$) defined for $\tilde{E} \in \tilde{\mathcal{Q}}$. If for a single number λ , with $|\lambda| = 1$ and $(\arg \lambda)/2\pi$ irrational, μ is invariant under rotation through $\arg \lambda$, i.e. $\mu(\lambda\tilde{E}) = \mu(\tilde{E})$ for every $\tilde{E} \in \tilde{\mathcal{Q}}$, then $\mu(\tilde{E}) = \tilde{m}(\tilde{E})$.

PROOF. Let \tilde{A}_1 and \tilde{A}_2 be any two closed intervals (arcs) of the same length in \tilde{X} . Since the sequence $\{\lambda^n\}$ of powers of λ is everywhere dense in X , we may find a sequence $\{n_j\}$ of positive integers, so that

$$(5) \quad \lim_{j \rightarrow \infty} \lambda^{n_j} \tilde{A}_1 = \tilde{A}_2,^{10}$$

and consequently

$$(6) \quad \lim_{j \rightarrow \infty} \mu(\lambda^{n_j} \tilde{A}_1) = \mu(\tilde{A}_2).^{11}$$

Since $\mu(\lambda^{n_j} \tilde{A}_1) = \mu(\tilde{A}_2)$, we have proved that $\mu(\tilde{A}_1) = \mu(\tilde{A}_2)$. Thus $\mu(\tilde{A})$ is a function of the arc length of \tilde{A} , i.e. of $\tilde{m}(\tilde{A})$. This numerical function is clearly monotone and additive, hence proportional to $\tilde{m}(\tilde{A})$. Considering $\tilde{A} = \tilde{X}$ shows that the factor of proportionality is 1. Thus $\mu(\tilde{E})$ and $\tilde{m}(\tilde{E})$ agree for arcs, and therefore for all Borel sets.

As an immediate consequence of this lemma we observe that if for any Borel set \tilde{E}_0 we have $\tilde{E}_0 = \lambda\tilde{E}_0$, then $\tilde{m}(\tilde{E}_0) = 0$ or else $\tilde{m}(\tilde{E}_0) = 1$, for otherwise

$$\mu(\tilde{E}) = \tilde{m}(\tilde{E}\tilde{E}_0)/\tilde{m}(\tilde{E}_0)$$

would contradict what we just proved.

After these preliminaries we are now ready to introduce the pathological concept we mentioned at the beginning of this section.

DEFINITION 4. A (not necessarily measurable) subset E of a measure space X is absolutely invariant if for every measure preserving transformation T of X on itself, the symmetric difference $(E - TE) + (TE - E)$ is measurable and has measure zero.

LEMMA 9. If E is measurable and $m(E) = 0$ or $m(E) = m(X)$ then E is absolutely invariant. Conversely if X is separable and non-atomic and $E \subset X$

¹⁰ See S. Saks, *Theory of the integral*, Warszawa, 1937, p. 5.

¹¹ See Saks, *op. cit.*, p. 8.

is measurable and absolutely invariant, then $m(E) = 0$ or $m(E) = m(X)$; if X is not measurable and absolutely invariant then $m_*(E) = 0$, $m^*(E) = m(X)$.

PROOF. The first statement is obvious. To prove the remaining statements we observe that if T is a measure preserving transformation and if A is a set (almost) invariant under T , in the sense that $(A - TA) + (TA - A)$ is measurable and has measure zero, then any measurable cover, A^* , and any measurable kernel, A_* , of A are also (almost) invariant under T .¹² For $A \subset A^*$ implies $TA \subset TA^*$; since TA and A are almost equal, and T is measure preserving, TA^* is a measurable cover of TA , and therefore $TA^* + (A - TA)$ is a measurable cover of A . It follows (since any two measurable covers of A are almost equal) that TA^* is almost equal to A^* , as was to be proved. A similar argument applies to measurable kernels.

It follows from the preceding paragraph that if E is absolutely invariant then so are E_* and E^* . If we knew that a measurable absolutely invariant set must have measure zero or $m(X)$, we could conclude that for a non-measurable absolutely invariant E , $m_*(E) = 0$ and $m^*(E) = m(X)$. In the case where X is the perimeter of the unit circle, there are many examples of measure preserving transformations whose measurable invariant sets all have measure zero or $m(X)$: in fact the rotations described in Lemma 8 are such. If a set is invariant under all measure preserving transformations it is *a fortiori* invariant under these and hence if it is measurable it will have measure zero or $m(X)$. The general case is, however, reduced to the case of the circle by Theorem 1.

To show that the concept of absolute invariance is not vacuous we shall now show that non-measurable absolutely invariant sets exist. In the existence proof we make free use of the continuum hypothesis and well ordering.

LEMMA 10. If $X = X(\mathfrak{C}, \mathfrak{A}, m)$ is a proper measure space of total measure one, there exists an absolutely invariant set $E \subset X$ with $m_*(E) = 0$, $m^*(E) = 1$.

PROOF. Since on a separable measure space there are at most \mathfrak{c} (= the power of the continuum) set transformations (since a set transformation is completely determined by its behavior on a countable collection of sets, and the set of all functions from a set of power \aleph_0 to a set of power \mathfrak{c} has power \mathfrak{c}), it follows from lemma 5 that we may find a set of at most \mathfrak{c} measure preserving transformations of X on itself with the property that every measure preserving transformation differs on at most a set of measure zero from one of the given set. Let this set be well ordered, so that to each ordinal $\alpha < \Omega$ (= the first uncountable ordinal) there corresponds a measure preserving transformation T_α . We may similarly enumerate the collection of all Borel sets of positive measure: let these be denoted by E_α , $\alpha < \Omega$.

For any $x \in X$ and any $\alpha < \Omega$ we write

$$C_\alpha(x) = \left\{ \prod_{i=1}^k T_{\alpha_i}^{n_i} x \mid \alpha_i = \alpha, k = 1, 2, \dots; n_i = 0, \pm 1, \pm 2, \dots \right\}.$$

¹² A^* [or A_*] is a measurable cover [or kernel] of A if it is measurable, if $A \subset A^*$ [or $A_* \subset A$], and if $m^*(A) = m(A^*)$ [or $m_*(A) = m(A_*)$]. If A_1^* and A_2^* are measurable covers of A then $(A_1^* - A_2^*) + (A_2^* - A_1^*)$ has measure zero.

$C_\alpha(x)$ is the smallest set containing x and invariant under T_β for all $\beta \leq \alpha$. Further relevant properties of $C_\alpha(x)$ are the following. $C_\alpha(x)$ is a countable set; for $\alpha \leq \beta$, $C_\alpha(x) \subset C_\beta(x)$; if $y \notin C_\alpha(x)$, then $C_\alpha(y)$ and $C_\alpha(x)$ are disjoint.

By transfinite induction we now define points x_α and y_α . x_1 is chosen in E_1 ; y_1 is chosen in E_1 but not in $C_1(x_1)$. Since $C_1(x_1)$ is countable and E_1 (being a Borel set of positive measure) is not, the choice of y_1 is possible. If x_α and y_α are defined for all $\alpha < \beta$, we define x_β as follows. Since the set

$$\sum_{\alpha < \beta} \{C_\beta(x_\alpha) + C_\beta(y_\alpha)\}$$

is countable, we may choose $x_\beta \in E_\beta$ so that x_β is not in this set. After this is done we may add $C_\beta(x_\beta)$ to this set and choose y_β so that $y_\beta \in E_\beta$, but y_β is not in the enlarged set.

Concerning the points x_α and y_α we now assert: for any α and β , $\alpha \neq \beta$, $C_\alpha(x_\alpha)$ and $C_\beta(y_\beta)$ are disjoint. If $\alpha \leq \beta$, then we know, by definition, that $y_\beta \notin C_\beta(x_\alpha)$ so that $C_\beta(y_\beta)$ and $C_\beta(x_\alpha)$ are disjoint—*a fortiori* $C_\beta(y_\beta)$ and $C_\alpha(x_\alpha)$ are disjoint. If $\alpha > \beta$, then again x_α is not in $C_\alpha(y_\beta)$ so that $C_\alpha(x_\alpha)$ and $C_\alpha(y_\beta)$ are disjoint, and therefore so also are $C_\alpha(x_\alpha)$ and $C_\beta(y_\beta)$.

We write

$$A = \sum_{\alpha < \Omega} C_\alpha(x_\alpha);$$

$$B = \sum_{\beta < \Omega} C_\beta(y_\beta);$$

it follows that A and B are disjoint. Since A contains x_α and B contains y_β , both A and B have at least one point in common with every Borel set of positive measure; consequently $X - A$ and $X - B$ cannot contain any such sets. It follows that both A and B have outer measure one (since their complements have inner measure zero), and since each is contained in the complement of the other, they both have inner measure zero.

It is now easy to see that A is (almost) invariant under every measure preserving transformation T . Given T we may find $\beta < \Omega$, such that T and T_β differ on at most a set of measure zero. Also we have

$$T_\beta A = \sum_{\alpha < \Omega} T_\beta C_\alpha(x_\alpha).$$

Since for $\alpha \geq \beta$, $C_\alpha(x_\alpha)$ is invariant under T_β , A and $T_\beta A$ can differ at most on the countable set $\sum_{\alpha < \beta} T_\beta C_\alpha(x_\alpha)$. Since $T_\beta A$ and TA differ on at most a set of measure zero, we have proved that A and TA differ on at most a set of measure zero. We may choose either A or B for the E of Lemma 10.

The following two lemmas establish the connection between absolute invariance and the property of having sufficiently many measure preserving transformations.

LEMMA 11. *Let $X(\mathcal{X}, m)$ be a measure space of total measure one with sufficiently many measure preserving transformations, and let $X' \subset X$ be any subset of X with $m^*(X') = 1$. If X' is absolutely invariant, then the measure space $X'(\mathcal{X}', m')$ (defined in Lemma 6) has sufficiently many measure preserving transformations.*

PROOF. The correspondence $E \rightleftharpoons E' = X'E$ is a set isomorphism between $\mathfrak{B} = \mathfrak{B}(\mathfrak{X})$ and $\mathfrak{B}' = \mathfrak{B}(\mathfrak{X}')$. Through this isomorphism any set mapping of X' on itself (i.e. any set isomorphism of \mathfrak{B}' on itself) induces a set mapping of X on itself. Since X has, by hypothesis, sufficiently many measure preserving transformations, it follows that to any set mapping T' on X' there corresponds a measure preserving transformation T of X on itself, such that T induces the same set mapping of X as T' . Since X' is absolutely invariant, $(X' - TX') + (TX' - X')$ has measure zero; let N' be the smallest set invariant under T which contains this set of measure zero. We may redefine T on N' to be the identity; the resulting T leaves X' strictly invariant and may therefore be considered as a measure preserving transformation of X' on itself. It is clear that this measure preserving transformation induces the set isomorphism T' on X' and that, therefore, X' has sufficiently many measure preserving transformations.

LEMMA 12. Let $X(\mathfrak{X}, m)$ be a measure space of total measure one which has a separating sequence of measurable sets, and let $X' \subset X$ be any subset of X with $m^*(X') = 1$. If the measure space $X'(\mathfrak{X}', m')$ (defined in Lemma 6) has sufficiently many measure preserving transformations then X' is an absolutely invariant subset of X .

PROOF. We use the notation introduced in the proof of Lemma 11. Let T be any measure preserving transformation on X ; through the correspondence $E \rightleftharpoons E' = X'E$, T induces a set mapping T' on X' . Since X' has sufficiently many measure preserving transformations, the set mapping T' of X' is induced by some measure preserving transformation, say S , of X' on itself. We shall prove that for almost every point $x \in X'$, $Sx = Tx$.

For any set $E \in \mathfrak{X}$ we know that $S^{-1}E' = S^{-1}(X'E)$ and $X' \cdot T^{-1}E$ differ on at most a set of measure zero (since S and T induce the same set mapping on X'): we denote this set of measure zero by N_E , and we write N for the union of all N_E , where we allow E to run through a separating sequence. Let x be any point in $X' - N$; we assert that $Sx = Tx$. If this were not true, we could find a set E , belonging to the separating sequence used above, such that $Sx \in E$ and $Tx \notin E$. Since $x \in X'$, $Sx \in X'$, and therefore $x \in S^{-1}(X'E)$; since $Tx \notin E$, *a fortiori* $x \notin X' \cdot T^{-1}E$. It follows that $x \in N_E \subset N$; since this contradicts the choice of x , we must have $Sx = Tx$.

We have proved that T leaves almost every point of X' in X' : in other words X' is almost invariant under T . Since T was arbitrary, it follows that X' is absolutely invariant.

We conclude this section with an isomorphism theorem that makes clear the structure of measure spaces with sufficiently many measure preserving transformations.

THEOREM 3. A necessary and sufficient condition that a proper measure space of total measure one have sufficiently many measure preserving transformations is that it be point isomorphic to an absolutely invariant subset of the unit interval.

PROOF. Since the property of possessing sufficiently many measure preserving

transformations is invariant under point isomorphism, and since, by Lemma 11, an absolutely invariant set has this property, sufficiency is clear.

To prove necessity we first observe that the given measure space, $X(\mathcal{C}, \mathcal{A}, m)$ is set isomorphic with $\tilde{X}(\tilde{\mathcal{C}}, \tilde{\mathcal{A}}, \tilde{m})$ in virtue of Theorem 1. (It will be most convenient in this proof to think of \tilde{X} as the perimeter of the unit circle in the complex plane.) Consider on \tilde{X} the measure preserving transformation $\tilde{x} \rightarrow \lambda \tilde{x}$, where $\lambda \in \tilde{X}$ is a fixed number with $(\arg \lambda)/2\pi$ irrational. The set isomorphism between \tilde{X} and X makes correspond to this transformation on \tilde{X} a certain measure preserving transformation T on X . A set isomorphism may also be considered as a mapping of the characteristic functions of X on the characteristic functions of \tilde{X} : this mapping may be extended to all $L_2(\tilde{X})$ and thus generates an isomorphism between $L_2(\tilde{X})$ and $L_2(X)$. Let $\phi(x)$ be the correspondent on X of the function $\tilde{\phi}(\tilde{x}) \equiv \tilde{x}$ on \tilde{X} ; the function $\phi(x)$ has the following properties:

- (i) $|\phi(x)| \equiv 1$;
- (ii) $\phi(Tx) \equiv \lambda \phi(x)$;
- (iii) $\{\phi^n(x)\} = \{(\phi(x))^n\}$, $n = 0, \pm 1, \pm 2, \dots$, is a complete orthonormal set in $L_2(X)$.

(To be precise: since $\phi(x)$ is determined only up to a set of measure zero, properties (i) and (ii) need to be true only almost everywhere. It is clear, however, that by changing ϕ on a set of measure zero we may assume that (i) and (ii) are always true. We may also assume, and we find it convenient to do so, that $\phi(x)$ is a Baire function.)

We apply Lemma 7 to $\{\phi^n(x)\}$ to obtain a set N of measure zero with the property described there. By increasing N , if necessary, we may assume that N is invariant under T . We now omit the points of N from X : we shall show that the remainder (henceforth to be denoted by X again) is in one to one measure preserving correspondence with an absolutely invariant subset of \tilde{X} .

The function $x' = \phi(x)$ defines a mapping from X to \tilde{X} ; we know that this mapping is Borel measurable (i.e. that the inverse image of a set in $\tilde{\mathcal{A}}$ lies in \mathcal{A}), and we assert furthermore that it is univalent. For if we had $\phi(x) = \phi(y)$, then we should also have $\phi^n(x) = \phi^n(y)$ for all n , and this possibility is precisely what we eliminated when we threw away the set N .

The transformation T is carried by the mapping ϕ into some transformation T' of the range $\phi(X) = X' \subset \tilde{X}$ into itself; since

$$T'x' = \phi(T\phi^{-1}(x')) = \lambda x',$$

we see that X' is invariant under the rotation $\tilde{x} \rightarrow \lambda \tilde{x}$.

For every Borel set $\tilde{E} \subset \tilde{X}$ (i.e. $\tilde{E} \in \tilde{\mathcal{A}}$) we define $\mu(\tilde{E}) = m(\phi^{-1}(\tilde{E}))$. Since $\phi(T\phi^{-1}(\tilde{E})) = X' \cdot \lambda \tilde{E}$, we have $T\phi^{-1}(\tilde{E}) = \phi^{-1}(X' \cdot \lambda \tilde{E}) = \phi^{-1}(\lambda \tilde{E})$. Since T is measure preserving it follows that

$$\mu(\tilde{E}) = m(\phi^{-1}(\tilde{E})) = m(T\phi^{-1}(\tilde{E})) = m(\phi^{-1}(\lambda \tilde{E})) = \mu(\lambda \tilde{E}).$$

Hence, by Lemma 8, $\mu(\tilde{E}) = \tilde{m}(\tilde{E})$.

Suppose, finally, that \tilde{E}_1 and \tilde{E}_2 are Borel subsets of \tilde{X} for which $X'\tilde{E}_1 = X'\tilde{E}_2$. Write $\tilde{E} = (\tilde{E}_1 - \tilde{E}_2) + (\tilde{E}_2 - \tilde{E}_1)$; it follows that $X'\tilde{E}$ is empty, so that $\phi^{-1}(\tilde{E})$ is empty and $\mu(\tilde{E}) = m(\phi^{-1}(\tilde{E})) = \tilde{m}(\tilde{E}) = 0$. This implies that $\tilde{m}(\tilde{E}_1) = \tilde{m}(\tilde{E}_2)$; it follows from Lemma 6 that $\tilde{m}^*(X') = 1$.

To sum up: we have proved that X is point isomorphic with a possibly non-measurable subset X' of \tilde{X} , with $\tilde{m}^*(X') = 1$; since X has sufficiently many measure preserving transformations, so does X' . Lemma 12 now applies: X' is absolutely invariant and the theorem is proved.

4. Application of the geometric isomorphism theorem to measure preserving transformations

In this section we shall have occasion to use certain facts about measure preserving transformations and the Pontrjagin duality theory: we describe briefly the parts of these theories that we need. Throughout the remainder of our work we consider only normal spaces of total measure one.

Two measure preserving transformations T_1 and T_2 , defined, say, on X_1 and X_2 , are (point —) *isomorphic*¹³ if there is a point isomorphism T from X_1 to X_2 with the property that TT_1T^{-1} is almost everywhere equal to T_2 . With every measure preserving transformation T we associate a unitary transformation U defined on $L_2(X)$ by $Uf(x) = f(Tx)$. A measure preserving transformation T has *pure point spectrum* if U has; in other words if there exists a complete orthonormal sequence, $\{f_n(x)\}$ of functions in $L_2(X)$ and a sequence $\Lambda = \{\lambda_n\}$ of complex numbers (of absolute value one) such that $f_n(Tx) = \lambda_n f_n(x)$ almost everywhere, for $n = 1, 2, \dots$. T is *ergodic* if $f(Tx) = f(x)$ almost everywhere, with $f \in L_2(X)$, is equivalent to $f(x) = \text{constant}$ almost everywhere. The *spectrum*, Λ , of an ergodic measure preserving transformation with pure point spectrum is a subgroup of the multiplicative group of complex numbers of absolute value one. The numbers $\lambda_n \in \Lambda$ are, moreover, a complete set of invariants of T , in the sense that if two measure preserving transformations with pure point spectrum have the same set Λ of eigenvalues with the same multiplicities then they are isomorphic.¹⁴

Concerning groups we shall need the following. A compact abelian separable topological group, X , as an m -space, in the sense that we may define on it an invariant metric $d(x, y)$ and (unique) invariant Haar measure $m(E)$ in such a way that it becomes an m -space.¹⁵ Let Λ' be the character group of X ; i.e. Λ' is the set of all complex valued continuous functions $f(x)$ with $|f(x)| \equiv 1$

¹³ Since this is the only kind of isomorphism for measure preserving transformations that we shall use, we shall in the sequel omit the qualifying 'point —'.

¹⁴ All these statements are proved in (I) for flows: it is easy, however, to make the translation from the one parametric case to the discrete case.

¹⁵ Invariance means that for all points x, y , and a , and all measurable sets E , we have $d(x, y) = d(ax, ay)$ and $m(aE) = m(E)$. We find it convenient to write all groups multiplicatively, even though they are abelian.

and $f(xy) = f(x)f(y)$. Then Λ' is countable, and the functions $f(x) \in \Lambda'$ form a complete orthonormal set in $L_2(X)$. Conversely let Λ be any countable abelian group, and let X be its character group; i.e. X is the set of all complex valued functions $x(\lambda)$, defined on Λ , with $|x(\lambda)| \equiv 1$ and $x(\lambda\mu) = x(\lambda)x(\mu)$. X may be so topologized that it becomes a compact separable (and, of course, abelian) group. If to every $\lambda \in \Lambda$ we make correspond the function $f(x)$ on X , defined by $f(x) = x(\lambda)$ then this correspondence is an isomorphism between Λ and the entire character group Λ' of X .¹⁶

The fact that Haar measure is invariant means that the rotation $x \rightarrow ax$, where a is any fixed element of the group, is a measure preserving transformation. The point of introducing the seemingly irrelevant compact groups into the study of measure preserving transformations is that such rotations are normal forms for a large class of transformations.

THEOREM 4. *An ergodic measure preserving transformation with pure point spectrum on a normal space is isomorphic to a rotation on a compact separable abelian group.*

PROOF. Let Λ be the spectrum of the given measure preserving transformation; let X be the character group of Λ , and Λ' that of X . If for every $\lambda \in \Lambda$ we define $a(\lambda) \equiv \lambda$, then $a = a(\lambda)$ is in X . For every $x \in X$ we define $Tx = ax$; we assert that T has pure point spectrum and that its spectrum is simple and precisely equal to Λ . It has pure point spectrum because the characters $f(x) \in \Lambda'$ form a complete orthonormal system on $L_2(X)$, and every such f is an eigenfunction of T belonging to the eigenvalue $f(a)$, $f(ax) = f(a)f(x)$. This shows, moreover, that the spectrum of T , including multiplicities, is obtained by forming the numbers $f(a)$ for all $f \in \Lambda'$. Since to each f there corresponds (through the isomorphism described above) an element $\lambda \in \Lambda$ for which $f(x) = x(\lambda)$ for all x , we see that we may equally well form the numbers $a(\lambda)$, i.e. λ , for all $\lambda \in \Lambda$. Hence T is ergodic and it follows, from the previously quoted result of (I), that the given transformation and T are isomorphic.

Since in this proof we used only the group Λ of eigenvalues and not the actual transformation we have also the following corollary.

COROLLARY 1. *Every countable group of complex numbers of absolute value one is the spectrum of an ergodic measure preserving transformation with pure point spectrum.*

Theorem 4 also enables us to characterize the set of all transformations which commute with a given ergodic transformation with pure point spectrum. The solution of this problem for general measure preserving transformations is probably very difficult.

COROLLARY 2. *If $x \rightarrow ax = Tx$ is an ergodic rotation on a compact abelian group X and if S is any measure preserving transformation on X for which $ST = TS$ then S is also a rotation.*

¹⁶ For the proof of all these statements see L. Pontrjagin, *Topological groups*, Princeton, 1939, Chapter V.

PROOF. We have $S(ax) = aS(x)$, so that if we write $b(x) = Sx \cdot x^{-1}$, then

$$b(Tx) = S(ax)(ax)^{-1} = Sx \cdot x^{-1} = b(x).$$

In other words $b(x)$ is invariant under T ; since T is ergodic $b(x) = b = \text{constant}$ ¹⁷ and $Sx = bx$, as was to be proved.

We shall call a measure preserving transformation R an *involution* if $R^2 = I$ (= the identity), and we shall call an involution a *factor* of a given transformation T if $S = RT$ is also an involution (so that $T = SR$).

COROLLARY 3. If $x \rightarrow ax = Tx$ is any rotation on a compact abelian group X , then T may be factored, $T = SR$, $S^2 = R^2 = I$; if T is ergodic every factor R of T is a reflection, $Rx = bx^{-1}$.

PROOF. Clearly if $Rx = bx^{-1}$ then R is an involution; also $Sx = RTx = R(ax) = ba^{-1} \cdot x^{-1}$ is an involution. Conversely if T is ergodic and if $T = SR$, $S^2 = R^2 = I$, then $TRT = SR \cdot R \cdot SR = R$, so that $aR(ax) = Rx$. It follows as in the proof of Corollary 2 that $b(x) = x \cdot R(x)$ is invariant under T , (i.e. $b(ax) = ax \cdot R(ax) = x \cdot R(x) = b(x)$), so that $b(x) = b = \text{constant}$, and $Rx = bx^{-1}$.

COROLLARY 4. Any ergodic measure preserving transformation T with pure point spectrum is isomorphic to its own inverse, $T^{-1} = RTR^{-1}$, where R may even be chosen as an involution.

PROOF. From Corollary 3 we know that $T = SR$, $S^2 = R^2 = I$; since $T^{-1} = R^{-1}S^{-1} = RS$, we have $T^{-1} = R \cdot SR \cdot R = R \cdot T \cdot R^{-1}$.

There seems to be some reason for the conjecture that the results of Corollaries 3 and 4 are valid for an arbitrary measure preserving transformation.

We have seen that every rotation is a measure preserving transformation with pure point spectrum; the question arises as to when a rotation is ergodic. The following theorem asserts that for rotations ergodicity (i.e. metric transitivity) is equivalent to regional transitivity.¹⁸

THEOREM 5. If a is a fixed element of the compact abelian group X , the rotation $x \rightarrow ax$ is ergodic if and only if the sequence $\{a^n\}$ is everywhere dense in X .

PROOF. If $x \rightarrow ax$ is ergodic then the iterates of some point, say x_0 , are everywhere dense.¹⁹ Since the transformation $x \rightarrow x \cdot x_0^{-1}$ is a homeomorphism, it carries the sequence $\{a^n x_0\}$ of iterates of x_0 into a dense sequence; but $a^n x_0 x_0^{-1} = a^n$.

Suppose, conversely, that $\{a^n\}$ is everywhere dense. We have already seen that any rotation has every function f in the character group Λ' of X for an eigenfunction, and that the functions of Λ' are a complete orthonormal set in $L_2(X)$. Since eigenfunctions belonging to different eigenvalues are orthogonal, every function invariant under the rotation $x \rightarrow ax$ must be a linear combina-

¹⁷ The definition of ergodicity says that numerically valued invariant functions are constant. It is easy to verify that this implies the same result for functions (such as $b(x)$) whose values are in the group X .

¹⁸ For a discussion of the various kinds of transitivity see G. A. Hedlund, *The dynamics of geodesic flows*, Bulletin of the American Mathematical Society, vol. 45, (1939), p. 243.

¹⁹ See Eberhard Hopf, *Ergodentheorie*, Berlin, 1937, p. 29.

tion of the invariant functions of the set Λ' : if the only invariant function in Λ' is $f(x) \equiv 1$, the rotation is ergodic. Suppose then that $f(ax) = f(x)$ for some $f \in \Lambda'$. It follows (taking x to be the unit element of X , $x = 1$) that $f(a^n) = f(1) = 1$ for all n ; since $\{a^n\}$ is dense and f is continuous it follows that $f(x) = 1$.

To introduce the final result of this paper we observe that Theorem 4, and the existence of an invariant metric on any compact separable group, imply that every ergodic measure preserving transformation with pure point spectrum is isomorphic to an isometric transformation on an m -space. Conversely:

THEOREM 6. *If T is an ergodic measure preserving transformation on an m -space $X(\mathcal{X}, m)$ such that to every $\epsilon > 0$ there corresponds a $\delta = \delta(\epsilon) > 0$ in such a way that $d(x, y) < \delta$ implies $d(T^n x, T^n y) < \epsilon$, $n = 0, \pm 1, \pm 2, \dots$, (in other words if the family $\{T^n\}$ of transformations is equicontinuous), then T has pure point spectrum: in fact it is possible to introduce into X a multiplication so that it becomes (with the original topology of X) a compact separable abelian group and T becomes a rotation.*

We comment first of all on the hypothesis. Since an isometric transformation clearly has the described equicontinuity property, on the face of it our hypothesis is weaker than isometry. But if our hypothesis is satisfied we may introduce into X a new metric, $d'(x, y)$, defined by

$$d'(x, y) = \sup \{ \min(1, d(T^n x, T^n y)) \mid n = 0, \pm 1, \pm 2, \dots \};$$

it is easy to verify that d and d' induce the same topology on X , and that $d'(Tx, Ty) = d'(x, y)$. We may (and do) therefore assume that T is isometric in the first place.

We shall make the proof of Theorem 6 depend on the following two lemmas which have an interest of their own.

LEMMA 13. *If on an m -space X there exists an ergodic and isometric measure preserving transformation then X is compact.*

PROOF. Let T be an ergodic and isometric transformation; since X is complete we have to show only that it is totally bounded. If it is not, then there is an $\epsilon > 0$ and an infinite sequence of points x_1, x_2, \dots in X such that the open spheres S_n of radius ϵ with center at x_n are pairwise disjoint. Let x_0 be any point of X whose iterates $\{T^k x_0\}$ are everywhere dense in X , and choose for each $n = 1, 2, \dots$ an integer $k = k(n)$ such that $d(x_n, T^k x_0) < \epsilon/2$. If we denote by S_0 the open sphere of radius $\epsilon/2$ with center at x_0 , then for each n , $T^{k(n)} S_0 \subset S_n$, so that $m(S_n) \geq m(S_0) > 0$. Since a measure space has, by definition, finite measure, there cannot exist an infinite sequence of pairwise disjoint sets whose measure is bounded away from zero; it follows that X is totally bounded and therefore compact.

LEMMA 14. *Let X be any compact group (not necessarily separable or abelian) and let $m(E)$ be any finite measure, defined (at least) for all Borel sets of X , such that the measure of an open set is positive and that the measure of any measurable set is the lower bound of the measure of open sets containing it. Then the set X_0 of all $x \in X$ for which $m(xE) = m(E)$ for all measurable sets E is a closed subgroup of X .*

PROOF. Since $x \in X_0$ and $y \in X_0$ implies

$$m(xy^{-1}E) = m(y^{-1}E) = m(y(y^{-1}E)) = m(E),$$

X_0 , and consequently its closure \bar{X}_0 , is a subgroup; we shall prove $\bar{X}_0 \subset X_0$.

Take $x \in \bar{X}_0$, and let E be any closed (and hence compact) subset of X . Let O be any open set, $O \supset xE$, and let N be a neighborhood of 1 (= the unit element of X) such that for $a \in N$, $axE \subset O$. Then Nx is a neighborhood of x , so that the intersection of Nx and X_0 is not empty; say $y = ax$, $a \in N$, $y \in X_0$. Then

$$m(E) = m(yE) = m(axE),$$

and since $axE \subset O$, $m(E) \leq m(O)$. In other words $xE \subset O$ implies that $m(E) \leq m(O)$; our condition on m implies that $m(E) \leq m(xE)$. Applying this result to the compact set xE and the point $x^{-1} \in \bar{X}_0$ (in place of E and x) we obtain $m(xE) \leq m(E)$, so that $m(xE) = m(E)$ for all closed sets E . It follows that $m(xE) = m(E)$ for all measurable sets E , as was to be proved.

PROOF OF THEOREM 6. Let x_0 be any point in X for which $\{T^n x_0\}$ is everywhere dense; write $x_n = T^n x_0$ for $n = \pm 1, \pm 2, \dots$. For $x = x_n$ and $y = x_m$ we define $p(x, y) = x_{n+m}$, and $r(x) = x_{-n}$. If $x' = x_{n'}$, $x'' = x_{n''}$, $y' = x_{m'}$, $y'' = x_{m''}$, then

$$\begin{aligned} d(p(x', y'), p(x'', y'')) &= d(x_{n'+m'}, x_{n''+m''}) \\ &\leq d(x_{n'+m'}, x_{n'+m''}) + d(x_{n'+m''}, x_{n''+m''}) \\ &= d(x_{m'}, x_{m''}) + d(x_{n'}, x_{n''}) \\ &= d(y', y'') + d(x', x''); \end{aligned}$$

in other words $p(x, y)$ is uniformly continuous throughout its domain of definition; similarly since we have

$$d(r(x), r(y)) = d(x_{-n}, x_{-m}) = d(x_{-n+n+m}, x_{-m+n+m}) = d(y, x),$$

$r(x)$ is uniformly continuous throughout its domain. The domain of $p(x, y)$ is an everywhere dense subset of the product space of X with itself, and the domain of $r(x)$ is an everywhere dense subset of X , consequently they each have a unique continuous extension, to all the product space and all X respectively.

The rest of the proof is now easy. We define, for every x and y in X , $xy = p(x, y)$ and $x^{-1} = r(x)$; it is clear that with these definitions X becomes an abelian topological group. We may write, for any $x = x_n$ and an arbitrary y , $p'(x, y) = T^n y$; then $p'(x, y)$ is a continuous extension of our original $p(x, y)$ and therefore (because of the uniqueness of extension) $T^n y = x_n y$. (For $n = 1$, we obtain, in particular, $Ty = x_1 y$ for all y . The originally chosen element x_0 is now the unit element of the group.) If E is any measurable set then $T^n E = x_n E$ has the same measure as E , so that measure is preserved by an everywhere dense set of x 's; since, by Lemma 13, X is compact, Lemma 14 implies that for all x and all measurable sets E , $m(xE) = m(E)$. The uniqueness of Haar measure implies that m is the Haar measure of the group X ; this completes the proof that T is a rotation, and hence has pure point spectrum.

THE BROWNIAN MOVEMENT AND STOCHASTIC EQUATIONS

By J. L. DOOB

(Received January 14, 1942)

The irregular movements of small particles immersed in a liquid, caused by the impacts of the molecules of the liquid, were described by Brown in 1828.¹ Since 1905 the Brownian movement has been treated statistically, on the basis of the fundamental work of Einstein and Smoluchowski. Let $x(t)$ be the x -coordinate of a particle at time t . Einstein and Smoluchowski treated $x(t)$ as a chance variable. They found the distribution of $x(t) - x(0)$ to be Gaussian, with mean 0 and variance $\alpha |t|$, where α is a positive constant which can be calculated from the physical characteristics of the moving particles and the given liquid. More exactly, such a family of chance variables $\{x(t)\}$ is now described as the family of chance variables determining a temporally homogeneous differential stochastic process: the distribution of $x(s + t) - x(t)$ is Gaussian, with mean 0, variance $\alpha |t|$, and if $t_1 < \dots < t_n$,

$$x(t_2) - x(t_1), \dots, x(t_n) - x(t_{n-1})$$

are mutually independent chance variables. Wiener, who was the first to discuss this stochastic process rigorously, proved in 1923 that the functions $x(t)$ of this stochastic process are continuous, with probability 1.² This is of course a desirable result, which makes the stochastic process somewhat more acceptable as the mathematical idealization of the Brownian movement. It was not expected, however, that the above distribution of $x(s + t) - x(s)$ would prove correct for small t . Even if the derivation did not break down for small t , the mathematical fact that $x(s + t) - x(s)$ has standard deviation $\alpha |t|$ so that $x(s + t) - x(s)$ is of the order of magnitude of $|t|^{\frac{1}{2}}$, implying that $dx(s)/ds$ cannot be finite, would suggest the desirability of modifications of the Einstein-Smoluchowski distributions. In fact it is easily seen that (with probability 1) $x(t)$ is not even of bounded variation, so that the path curves of the Einstein-Smoluchowski process have infinite length!

A different stochastic process describing the $x(t)$ was in fact derived in 1930 by Ornstein and Uhlenbeck (15),³ and later by S. Bernstein (1), (2) and Krutkow (11), all using different methods. This new distribution of $x(s + t) - x(s)$ is

¹ For a historical account of the subject up to 1913, see Haas-Lorentz (6). (The numbered references will refer to the bibliography at the end of the paper.)

² Wiener (18, pp. 148-151) has since given a more simple proof. For a discussion of the exact meaning of such a statement concerning the continuity of paths, cf. Doob (3) and (5), §2. The result means that $x(t)$ can be treated as representing one of a multiplicity of continuous functions of t , and integrated, etc. Probability here is formally the study of measures on certain spaces of functions.

³ Cf. also Ornstein and Wijk (16) and Wijk (17). References to work since 1913 are given in Ornstein and Uhlenbeck (15).

Gaussian, with mean 0 and variance $(\alpha/\beta)(e^{-\beta|t|} - 1 + \beta|t|)$, approximately $\alpha|t|$ for large t , but $\alpha\beta t^2/2$ for small t . (Here β is a second physically determined constant.)

The purpose of the present paper is to apply the methods and results of modern probability theory to the analysis of the Ornstein-Uhlenbeck distribution, its properties and its derivation. It will be seen that the use of rigorous methods actually simplifies some of the formal work, besides clarifying the hypotheses. A stochastic differential equation will be introduced in a rigorous way to give a precise meaning to the Langevin differential equation for the velocity function $dx(s)/ds$. This will avoid the usual embarrassing situation in which the Langevin equation, involving the second derivative of $x(s)$ is used to find a solution $x(s)$ not having a second derivative.

1. The velocity distribution

The displacement function $x(t)$, as discussed by Ornstein and Uhlenbeck, has a derivative $u(t)$, and all the probability relations needed can be derived from those of $u(t)$, as will be seen below. The distribution of $u(t)$ can be described as follows: the conditional distribution of $u(s+t)$ ($t > 0$) for given $u(s) = u_0$, is Gaussian, with mean $u_0 e^{-\beta t}$ and variance $\sigma_0^2(1 - e^{-2\beta t})$. Here σ_0^2 , β are physically determined constants. When $t \rightarrow \infty$, this distribution becomes the Maxwell distribution of velocities, furnishing stationary absolute (unconditioned) probabilities for the process, if these are desired. Using these absolute probabilities, which make the distribution easier to describe, the full description of the $u(t)$ distribution can then be stated as follows: for each t , $u(t)$ is a chance variable with a Gaussian distribution, having mean 0, variance σ_0^2 ; the transition probabilities are as just described; the process is a Markoff process.⁴ This last fact means that the Maxwell distribution of $u(t_0)$ for each fixed t_0 , and the transition probabilities just described determine the full set of probability relations of the process. Under these conditions, if $t_1 < t_2$, the pair $u(t_1)$, $u(t_2)$ has a bivariate Gaussian distribution, with zero means, equal variances σ_0^2 , and correlation coefficient $e^{-\beta(t_2-t_1)}$. This stochastic process goes back at least to Smoluchowski, although it was first derived by Ornstein and Uhlenbeck as the process describing the velocity of a particle in Brownian motion. Ornstein and Uhlenbeck were only interested in the transition probabilities. The formal manipulations made below will show that there are technical advantages in defining (unconditioned) probabilities for the $u(t)$ also. The above described process will be called the O. U. process below.

The following theorem shows that such a process is essentially determined by three fundamental properties, of which at least the first two have simple physical

⁴ A process is called a Markoff process if whenever $t_1 < \dots < t_n$, the conditional distribution of $u(t_n)$ for given values of $u(t_1), \dots, u(t_{n-1})$ actually depends only on $u(t_{n-1})$. It is in this case, and only in this case, that the Smoluchowski equation between the transition probabilities, and the Fokker-Planck differential equations for the transitional probabilities are valid.

significance. (We can exclude Case A of the theorem, since it obviously does not fit the physical picture.)

THEOREM 1.1. *Let $u(t)$ ($-\infty < t < +\infty$) be a one-parameter family of chance variables, determining a stochastic process with the following properties.*

1. *The process is temporally homogeneous.*⁵

2. *The process is a Markoff process.*

3. *If s, t are arbitrary distinct numbers, $u(s), u(t)$ have a (non-singular) bivariate Gaussian distribution.*

Define m, σ_0^2 by

$$(1.1.1) \quad m = E\{u(t)\}, \quad \sigma_0^2 = E\{[u(t) - m]^2\}.$$

Then the given process is one of the following two types.

(A) *If $t_1 < \dots < t_n$, $u(t_1), \dots, u(t_n)$ are mutually independent Gaussian chance variables, with mean m and variance σ_0^2 .*

(B) *(O. U. process) There is a constant $\beta > 0$ such that if $t_1 < \dots < t_n$, $u(t_1), \dots, u(t_n)$ have an n -variate Gaussian distribution, with common mean m and variance σ_0^2 , and correlation coefficients determined by the equation $E\{[u(t) - m][u(s) - m]\} = \sigma_0^2 e^{-\beta|t-s|}$.*

Instead of considering $u(t)$, we can consider $(1/\sigma_0)[u(t) - m]$, which has mean 0 and variance 1. Then we shall assume in the following that $u(t)$ itself has these properties: $m = 0$, $\sigma_0^2 = 1$. Let $\rho(t)$ be the correlation function: $\rho(t) = E\{u(s+t)u(s)\}$, independent of s by Property 1. If $s < t$, the conditional distribution of $u(t)$ for given $u(s)$ has density

$$(1.1.2) \quad \frac{1}{(2\pi)^{\frac{1}{2}}(1-\rho^2)^{\frac{1}{2}}} \exp\left(-\frac{1}{2} \frac{[u(t) - \rho u(s)]^2}{1-\rho^2}\right), \quad \rho = \rho(t-s),$$

(Property 3). If $t_1 < \dots < t_n$, $u(t_1), \dots, u(t_n)$ then have an n -variate Gaussian distribution with density

$$(1.1.3) \quad \frac{1}{(2\pi)^{\frac{n}{2}} \prod_{i=1}^{n-1} (1-\rho_i^2)^{\frac{1}{2}}} \exp\left(-\frac{1}{2} u_1^2 - \frac{1}{2} \sum_{i=1}^{n-1} \frac{(u_{i+1} - \rho_i u_i)^2}{1-\rho_i^2}\right),$$

$$\rho_i = \rho(t_{i+1} - t_i), \quad u_i = u(t_i)$$

using Property 2. Now if u_1, \dots, u_n have an n -variate Gaussian distribution with density

$$(1.1.4) \quad \frac{1}{\Delta} \exp\left(-\frac{1}{2} \sum_{i,j} a_{ij} u_i u_j\right),$$

$\Delta = \det(E\{u_i u_j\})$ is the determinant of the matrix of variances and covariances, and (a_{ij}) is the inverse of this matrix. Using these facts we can calculate $\rho(t_3 - t_1) = E\{u_1 u_3\}$ in (1.1.3) with $n = 3$, and find that $\rho(t_3 - t_1) = \rho_1 \rho_2$, that is

⁵ That is, the probability distributions are unaffected by translations of the t -axis.

⁶ The expectation of the chance variable v will be denoted by $E\{v\}$.

$$(1.1.5) \quad \rho(t_3 - t_1) = \rho(t_2 - t_1)\rho(t_3 - t_2).$$

Then $\rho(t)$ is an even function; $|\rho(t)| \leq 1$ (Schwarz's inequality); and according to (1.1.5) $\rho(s+t) = \rho(s)\rho(t)$ for all positive s, t . Under these conditions either $\rho(t) \equiv 0$ or there is a constant $\beta \geq 0$ such that

$$(1.1.6) \quad \rho(t) = e^{-\beta|t|}.$$

In the present case, $\beta > 0$, by Property 3 (non-singularity of the given bivariate distributions). Evidently $\rho(t) \equiv 0$ furnishes Case A of the theorem, which certainly has the three given properties. If $\rho(t)$ is given by (1.1.6) with $\beta > 0$, we show first that the matrix (a_{ij}) , the inverse of $(\rho(t_j - t_i))$ actually determines a Gaussian density distribution (1.1.4). To see this we consider the density function (1.1.3) with $\rho_j = e^{-\beta(t_{j+1} - t_j)}$. The coefficients of the quadratic form in the exponent of (1.1.3) are easily evaluated and the matrix of the form is found to be the inverse of the matrix $(e^{-\beta|t_j - t_i|})$. Thus (1.1.3) actually is the required probability density. Moreover the probability densities obtained in this way (as the t_i vary) are mutually consistent, because integrating out any variable leaves a quadratic form of the same type, without the integrated variable, but with the same rule determining the coefficients. The correlation function (1.1.6) therefore determines a stochastic process. The process obviously is a Markoff process because of the form of the probability density (1.1.3): an initial factor involving u_1 only, followed by the product of functions of pairs of adjacent variables. The proof of the theorem is now complete.

According to a theorem of Khintchine ((9) p. 608), $\rho(t)$ is the correlation function of a temporally homogeneous stochastic process if and only if it can be put in the form

$$(1.1.7) \quad \rho(t) = \int_0^\infty \cos \lambda t \, dF(\lambda),$$

where $F(\lambda)$ is monotone non-decreasing and bounded. In Case B of the theorem, (1.1.7) is true when $F(\lambda)$ is given by

$$(1.1.8) \quad F(\lambda) = \frac{2\beta\sigma_0^2}{\pi} \int_0^\lambda \frac{d\lambda}{\beta^2 + \lambda^2}.$$

In the stochastic process of Case B, the variance of $u(s+t) - u(s)$ is $2\sigma_0^2\beta|t|$ for small t :

$$(1.1.9) \quad E\{[u(s+t) - u(s)]^2\} = 2\sigma_0^2(1 - e^{-\beta|t|}) \sim 2\sigma_0^2\beta|t|.$$

Thus $u(s+t) - u(s)$ is of the order of magnitude of $|t|^{1/2}$, and du/dt cannot exist. Physically this means that the particles in question do not have a finite acceleration (if the given stochastic process represents the Brownian movement that closely).

THEOREM 1.2. *If $u(t)$ is the representative function of the stochastic process of Theorem 1.1 Case B, $u(t)$ is a continuous function of t , with probability 1.*

Let $v(t)$ be determined by the equation

$$(1.2.1) \quad v(t) = t^{\frac{1}{2}} u \left(\frac{1}{2\beta} \log t \right), \quad t > 0.$$

Then $v(t)$ has the property that if $t_1 < \dots < t_n$, $v(t_1), \dots, v(t_n)$ have an n -variate Gaussian distribution. We find by direct calculation (taking $m = 0$):

$$(1.2.2) \quad \begin{aligned} E\{v(s+t) - v(s)\} &= 0, \\ E\{[v(s+t) - v(s)]^2\} &= \sigma_0^2 t, \\ E\{[v(s_2) - v(s_1)][v(t_2) - v(t_1)]\} &= 0, \quad (s_1 < s_2 \leq t_1 < t_2). \end{aligned}$$

Then $v(t)$ determines a differential process—in fact precisely the original Einstein-Smoluchowski process. Since Wiener has proved continuity of the path functions in this case, the theorem follows.

The transition from $u(t)$ to $v(t)$ just used reduces every property of the Ornstein-Uhlenbeck stochastic process to a corresponding property of the Einstein-Smoluchowski process, and vice versa. Many properties of the individual functions of the latter process, that is, properties possessed by almost all the individual functions, in other words possessed “with probability 1,” have been proved in recent years, besides the continuity property we have just used. The following theorem gives the counterparts of two of these for the O. U. process.

THEOREM 1.3. *If $u(t)$ is the representative function of the O. U. process of Theorem 1.1 Case B,*

$$(1.3.1) \quad \limsup_{t \rightarrow 0} \frac{u(t) - u(0)}{(4\sigma_0^2 \beta t \log \log (1/t))^{\frac{1}{2}}} = 1, \quad \limsup_{t \rightarrow 0} \frac{u(t)}{(2\sigma_0^2 \log t)^{\frac{1}{2}}} = 1,$$

with probability 1.

Let $v(t)$ be defined by (1.2.1). Then Khintchine ((10) pp. 68–75) has proved

$$(1.3.2) \quad \limsup_{t \rightarrow 0} \frac{v(1+t) - v(1)}{(2\sigma_0^2 t \log \log (1/t))^{\frac{1}{2}}} = 1, \quad \limsup_{t \rightarrow \infty} \frac{v(t) - v(0)}{(2\sigma_0^2 t \log \log t)^{\frac{1}{2}}} = 1,$$

and (1.3.2) becomes (1.3.1) when $v(t)$ is expressed in terms of $u(t)$.

2. The distribution of displacements

It does not seem to have been realized by earlier writers that the distribution of displacements in the O. U. process can be obtained directly from that of the velocities. In fact, we have seen that as t varies, $u(t)$ considered as one of a multiplicity of continuous functions of t . Integration of $u(t)$ is therefore admissible, and will give the displacement function. If $x(t)$ is the x -coordinate of a particle at time t ,

$$(2.1) \quad x(t) - x(0) = \int_0^t u(s) ds$$

with probability 1 (that is, neglecting the discontinuous $u(t)$ functions which have total probability 0). The main advantages of the rigorous approach to stochastic processes depending on a continuous parameter is precisely that the $u(t)$ of the process, as t varies, can be regarded as an individual function or rather, as one of many functions with whatever regularity properties the given probability distributions imply. Theorem 1.3 limits the actual upper bounds of the velocity functions $u(t)$. The following result takes advantage of the oscillations in sign.

THEOREM 2.1. *If $u(t)$ is the representative function of the O. U. process of Theorem 1.1 Case B, with $m = 0$,*

$$(2.1.1) \quad \lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t u(s) ds = \lim_{t \rightarrow \infty} \frac{x(t) - x(0)}{t} = 0,$$

with probability 1.

This theorem is simply the ergodic theorem applied to the $u(t)$ process to give the strong law of large numbers, (cf. Doob (4) p. 294). From (2.2.3) below, it is quite obvious that the expectation of the square of the left side of (2.1.1) goes to 0 as $t \rightarrow \infty$, so that the left side goes to 0 in the mean. The strength of (2.1.1) is that it is a statement about the path of the individual path functions, or physically, a statement about the path of a single particle. The same was true in Theorems 1.2 and 1.3.

In order to find the distribution of $x(t) - x(0)$ we proceed as follows. Riemann integrability of $u(t)$ implies that (with probability 1)

$$(2.2.1) \quad x(t) - x(0) = \lim_{n \rightarrow \infty} \sum_{j=1}^n u(tj/n) t/n.$$

Now the n -variate distribution of the variables summed is Gaussian. Then the sum is Gaussian, so the distribution of $x(t) - x(0)$ is also Gaussian, if it can be shown that the variance of $x(t) - x(0)$ is positive. The distribution of $x(t) - x(0)$ is thus completely determined by its first two moments, which we proceed to calculate. We shall suppose, that $E\{u(t)\} = 0$, $E\{u(t)^2\} = \sigma_0^2$. Then we find

$$(2.2.2) \quad E\{x(t) - x(0)\} = \int_0^t E\{u(s)\} ds = 0,$$

and, if $t > 0$,

$$(2.2.3) \quad \begin{aligned} E\{[x(t) - x(0)]^2\} &= \int_0^t \int_0^t E\{u(s)u(s')\} ds ds' \\ &= \sigma_0^2 \int_0^t \int_0^t e^{-\beta|s-s'|} ds ds' = \frac{2\sigma_0^2}{\beta^2} (e^{-\beta t} - 1 + \beta t). \end{aligned}$$

⁷ By Fubini's integration theorem, we can find the expectations under the integral sign, before integrating with respect to s .

The same sort of argument shows that if t_1, \dots, t_n are any distinct numbers, the chance variables

$$\{x(t_j) - x(0), u(t_j), \quad j = 1, \dots, n\}$$

have a $2n$ -variate Gaussian distribution, which can then be evaluated explicitly by finding the first and second moments. For example, the following equations determine the bivariate distribution of $x(t) - x(0), u(t), (t > 0)$:

$$(2.2.4) \quad \begin{aligned} E\{[x(t) - x(0)]u(t)\} &= \int_0^t E\{u(t)u(s)\} ds = \frac{\sigma_0^2}{\beta} (1 - e^{-\beta t}), \\ E\{x(t) - x(0)\} &= 0, \end{aligned}$$

$$E\{[x(t) - x(0)]^2\} = \frac{2\sigma_0^2}{\beta^2} (e^{-\beta t} - 1 + \beta t), \quad E\{u(t)\} = 0, \quad E\{u(t)^2\} = \sigma_0^2.$$

Thus the bivariate density of $x(t) - x(0), u(t)$ is Gaussian, with common mean 0, and variances $(2\sigma_0^2/\beta^2)(e^{-\beta t} - 1 + \beta t), \sigma_0^2$, respectively, and correlation coefficient

$$(2.2.5) \quad \frac{1 - e^{-\beta t}}{2(e^{-\beta t} - 1 + \beta t)^{1/2}}.$$

It is to be expected that if $s_1 < s_2 \leq t_1 < t_2$, $x(s_2) - x(s_1)$ and $x(t_2) - x(t_1)$ become independent as $t_1 \rightarrow \infty$. In fact, these two normally distributed variables have correlation coefficient

$$(2.2.6) \quad \frac{(e^{\beta s_2} - e^{\beta s_1})(e^{-\beta t_1} - e^{-\beta t_2})}{2(e^{-\beta(s_2-s_1)} - 1 + \beta(s_2-s_1))^{1/2}(e^{-\beta(t_2-t_1)} - 1 + \beta(t_2-t_1))^{1/2}};$$

which goes to 0 when t_1 and t_2 become infinite.

If in this discussion only the conditional distribution functions are wanted, for $u(0) = u_0$, for example, two procedures are possible. Setting $u(0) \equiv u_0$ instead of using the initial distribution we have used above, carrying out the same type calculations as above, now would give the desired conditional probabilities. Or the conditional distributions could be calculated from the distributions just derived, since the conditional distributions of a multivariate Gaussian distribution are easily found. Theorems 1.2, 1.3 and 2.1 hold no matter what initial distribution is assigned to $u(0)$.

Finally, there is one more fact which we shall need in the next section. Define $B(t)$ by

$$(2.2.7) \quad B(t) = \beta[x(t) - x(0)] + u(t) - u(0).$$

Then $B(t)$ has for each t a Gaussian distribution, with mean 0. Evidently the distribution of $B(s+t) - B(s)$ is independent of s . It is Gaussian, with mean 0, and the variance is easily calculated to be $2\sigma_0^2\beta|t|$. Moreover, if $s_1 < s_2 \leq t_1 < t_2$,

$$(2.2.8) \quad E\{[B(t_2) - B(t_1)][B(s_2) - B(s_1)]\} = 0.$$

Thus the $B(t)$ -process is again the Einstein-Smoluchowski process.

3. Derivation of the velocity distribution using the Langevin equation

Ornstein and Uhlenbeck base their investigation on the Langevin equation

$$(3.1) \quad \frac{du}{dt} = -\beta u(t) + A(t),$$

which is simply Newton's law of motion applied to a particle, after dividing through by the mass. The first term on the right is due to the frictional resistance or its analogue, which is supposed proportional to the velocity. The second term represents the random forces (molecular impacts). Probability hypotheses are imposed on the $A(t)$, including relations between $A(t)$ and $u(t)$, to determine the $u(t)$ distribution. Unfortunately this $u(t)$ distribution (Case B of Theorem 1.1), as we have seen, has the property that the velocity function has no time derivative. Then the solution can hardly satisfy (3.1).

Bernstein ((2) p. 361) replaces (3.1) by a finite difference equation:

$$(3.2) \quad \Delta \left(\frac{\Delta \xi_n}{\Delta t} \right) = -\beta \Delta \xi_n + \alpha_n, \quad n = 1, 2, \dots$$

Here ξ_1, ξ_2, \dots is a sequence of chance variables, $\Delta \xi_n = \xi_{n+1} - \xi_n$ etc., and $\alpha_1, \alpha_2, \dots$ is a given sequence of mutually independent chance variables. If we think of ξ_j as the analogue of $x(j\Delta t)$, the correspondence between (3.2) and (3.1) is clear. The equations of (3.2) determine definite distributions for the ξ_j in terms of those of the α_j . Bernstein shows that as $\Delta t \rightarrow 0$ the distribution of $\Delta \xi_n / \Delta t$ ($\sim \Delta x / \Delta t$) becomes the $u(t)$ distribution we have been discussing, if suitable hypotheses are made on the α_j . This approach is essentially different from that of Ornstein and Uhlenbeck in that Bernstein, as he states explicitly ((1) pp. 5, 6) is not writing a difference equation in the displacement functions $x(t)$ themselves: (3.2) determines distributions only, and these are approximated by the limiting distributions described in Theorem 1.1 Case B.

In our treatment, we shall replace the Langevin equation by a formalized differential equation for the velocity function $u(t)$. This equation is to be exact, not merely asymptotically true. The equation will be perfectly proper mathematically, so that solution by ordinary methods will provide all the information relevant to the desired distributions, and solution of more general problems, involving external forces, will require no special methods.

The problem is to find a proper stochastic analogue of the Langevin equation, remembering that we do not expect $u'(t)$ to exist. We write the equation in the following form:

$$(3.3) \quad du(t) = -\beta u(t) dt + dB(t),$$

and try to give these differentials a suitable interpretation. We shall suppose

that the $B(t)$ -process is a differential process: that is, if $t_1 < \dots < t_n$, we suppose that

$$B(t_2) - B(t_1), \dots, B(t_n) - B(t_{n-1})$$

are mutually independent chance variables. We also suppose temporal homogeneity, that is that the distribution of $B(s + t) - B(s)$ is independent of s . The physical meaning of these hypotheses is clear, and they will be justified further below. Equation (3.3) can be interpreted roughly in terms of small changes in momentum. An important particular case is that in which the second moments of the $B(t)$ -process are finite:

$$(3.4) \quad \sigma^2(t) = E\{[B(s + t) - B(s)]^2\} < \infty.$$

The first moment $E\{B(s + t) - B(s)\}$ then exists. If this first moment vanishes, $\sigma^2(t)$ satisfies the functional equation

$$\sigma^2(s + t) = \sigma^2(s) + \sigma^2(t).$$

Then $\sigma^2(t)$ must be proportional to t : $\sigma^2(t) = t\sigma^2$. If $f(t)$ is continuous,

$$(3.5) \quad \int_a^b f(t) dB(t)$$

has been defined under these hypotheses (Wiener (18), pp. 151–157, Doob (3); pp. 131–134), even though the functions $B(t)$ are known not to be of bounded variation. The definition makes all the formal processes correct. For example, if $f'(t)$ exists and is continuous,

$$(3.6) \quad \int_a^b f(t) dB(t) = f(t)[B(t) - B(0)] \Big|_a^b - \int_a^b [B(t) - B(0)]f'(t) dt^8$$

with probability 1. The usual Riemann-Stieltjes sums converge to (3.5) in the mean. Moreover

$$(3.7) \quad \begin{aligned} E \left\{ \int_a^b f(t) dB(t) \right\} &= 0, \\ E \left\{ \left[\int_a^b f(t) dB(t) \right] \left[\int_a^b g(t) dB(t) \right] \right\} &= \sigma^2 \int_a^b f(t)g(t) dt. \end{aligned}$$

Now it can be shown even without the hypothesis of the finiteness of the second moment in (3.4) that the formal integral in (3.5) can be defined, and will satisfy (3.6). The form of the characteristic function of $B(s + t) - B(s)$ has been derived by Lévy ((14) Chapter VII) and using this it is easy to prove that the

⁸ We never write $B(t)$ alone in an equation, since strictly speaking only differences like $B(t) - B(0)$ are defined. It is unnecessary to define $B(0)$ itself, although for convenience it can be taken identically 0, without affecting any of the equations to be used. Differential processes have been discussed in detail by Lévy ((12), (13), (14) Chapter VII) and Doob (3) §3).

usual Riemann-Stieltjes sums for the integral (3.5) converge in probability. The integral is defined as the limit, and (3.6) is readily verified. On the other hand, (3.7) cannot be expected to hold, since if $f(t) = 1$ the integral becomes $B(b) - B(a)$, and we have not supposed that the expectation of this difference is finite. The special case in which the second moment is finite is the only important one for the purposes of this section, but less restrictive conditions will be needed in §5. We shall justify later the assumption that the $B(t)$ process is a differential process.

We shall interpret an equation in differentials like (3.3) to mean the truth (with probability 1, that is for almost all functions $u(t)$) of

$$(3.8) \quad \int_a^b f(t) du(t) = -\beta \int_a^b f(t)u(t) dt + \int_a^b f(t) dB(t)$$

for all a, b , whenever f is a continuous function. Here the first two integrals are to be defined as the limits (in probability) of the usual Riemann or Riemann-Stieltjes sums. Equation (2.2.7) implies

$$(3.9) \quad \begin{aligned} \int_a^b f(t) du(t) &= -\beta \int_a^b f(t) dx(t) + \int_a^b f(t) dB(t) \\ &= -\beta \int_a^b f(t)u(t) dt + \int_a^b f(t) dB(t). \end{aligned}$$

Thus (3.3) holds for the $u(t)$ of the O. U. distribution if the $B(t)$ is defined by (2.2.7). Moreover (2.2.7) with $B(t)$ replaced by $B(t) - B(0)$ is an immediate consequence of (3.3). In this case, $B(t)$ has the property that the differences $B(s+t) - B(s)$ have finite second moments and, even Gaussian distributions, but we are not making either assumption in solving (3.3).

If (3.3) is true, then (with probability 1)

$$(3.10) \quad \int_0^t e^{\beta\tau} du(\tau) = -\beta \int_0^t e^{\beta\tau} u(\tau) d\tau + \int_0^t e^{\beta\tau} dB(\tau),$$

which implies, since integration by parts is applicable,

$$(3.11) \quad u(t) = u(0)e^{-\beta t} + e^{-\beta t} \int_0^t e^{\beta\tau} dB(\tau)$$

for all t , with probability 1. Conversely suppose that $u(t)$ is defined by (3.11). Since $B(t)$ is known to be continuous in t except for non-oscillatory discontinuities (jumps) (Lévy (12) pp. 359-364, (13); Doob (3), pp. 134-138), the same must be true of the right side of (3.11), and therefore of $u(t)$. Then $u(t)$ is Riemann integrable with probability 1. Moreover

$$(3.12) \quad \int_a^b f(t) e^{-\beta t} d_t \int_0^t e^{\beta\tau} dB(\tau) = \int_a^b f(t) dB(t),$$

so that from (3.11)

$$(3.13) \quad \int_a^b f(t) e^{-\beta t} d[e^{\beta t} u(t) - u(0)] = \int_a^b f(t) dB(t),$$

proving incidentally that the left side exists. The left side can be simplified to

$$(3.14) \quad \beta \int_a^b f(t) u(t) dt + \int_a^b f(t) du(t),$$

and putting this into (3.13) we find that (3.8) is satisfied. Then (3.11) furnishes the complete solution of (3.3) under the stated conditions. We stress again that although (3.11) implies strong connections between the $u(t)$ and $B(t)$ processes, we have made no such hypothesis in the derivation not implicit in (3.3). Lévy ((14) pp. 166–167) has shown that the only differential processes whose path functions $B(t) - B(0)$ do not have jumps have the property that the distribution of $B(t) - B(0)$ is Gaussian. Then it is only in this case, which will lead to the O. U. process, that $u(t)$ will not have jumps.

The term $\beta u(t)$ in the Langevin equation is supposed to account for the total frictional effect, including the Doppler friction, caused by the fact that more impacts decelerate than accelerate the motion of a moving particle. The term $A(t)$ in (3.1) or $dB(t)$ in (3.3) represents the “purely random” impulses, that is, the residual effect after the frictional effect has been subtracted out. One idea running through any treatment of the Langevin equation is that this term or, sometimes, $x(t)$ itself, is independent of the given velocity at any time. This hypothesis goes back to Langevin, and has caused much controversy. We shall make the hypothesis only to the following extent. The chance variable $u(0)$ will be given various initial distributions, but will always be made independent of the $B(t)$ -process for $t \geq 0$. This means that if $0 \leq t_1 < \cdots < t_n$ the chance variable $u(0)$ is supposed independent of the set of chance variables

$$\{B(t_{j+1}) - B(t_j), \quad j = 1, \cdots, n-1\}.$$

We shall describe the above hypothesis in the following physical terms: *the initial velocity $u(0)$ is independent of later residual random impacts.* It would be a serious drawback to the whole treatment if when $u(0)$ is so chosen $u(t_0)$ for each $t_0 > 0$ were not independent of the $B(t)$ -process for $t \geq t_0$, that is if $u(t_0)$ were not independent of later residual random impacts for all t_0 . We can prove, however, the following statement, which incidentally justifies our hypothesis that the $B(t)$ -process is a differential process. *Let the $B(t)$ process be a differential process, and define $u(t)$ by (3.11). If the chance variable $u(0)$ is independent of the $B(t)$ -process for $t \geq 0$, then $u(t_0)$ will be independent of the $B(t)$ -process for $t \geq t_0$, for all $t_0 > 0$. Conversely suppose only that the $B(t)$ -process is regular enough that the integral (3.5) can be defined as the limit in probability of the usual sums, and that (3.6) is true. Then if $u(t)$ is defined by (3.11), and if choosing $u(0)$ independent of the $B(t)$ process for $t \geq 0$ implies that $u(t_0)$ will be independent of the $B(t)$ -process for $t \geq t_0$, for all $t_0 > 0$, then the $B(t)$ -process is a differential process.*

PROOF. Let the $B(t)$ -process be a differential process, define $u(t)$ by (3.11) and let $u(0)$ be independent of the $B(t)$ -process for $t \geq 0$. Then from (3.11) with $t = t_0$, $u(t_0)$ involves only $u(0)$ and the $B(t)$ -process for $t \leq t_0$. Then $u(t_0)$ is independent of the $B(t)$ -process for $t \geq t_0$ because the $B(t)$ -process is a differential one, with differences involving t -values beyond t_0 independent of those involving t -values before t_0 . Conversely suppose that choosing $u(0)$ independent of the $B(t)$ -process for $t \geq 0$ implies that $u(t_0)$ will be independent of the $B(t)$ -process for $t \geq t_0$, for all $t_0 > 0$. Then if $u(0)$ is so chosen,

$$u(0) + \int_0^{t_0} e^{\beta\tau} dB(\tau)$$

and therefore

$$\int_0^{t_0} e^{\beta\tau} dB(\tau)$$

are independent of the $B(t)$ -process for $t \geq t_0$. This fact implies that the preceding integral determines a differential process, that is, if $t_1 < \dots < t_n$, the integrals

$$\int_{t_j}^{t_{j+1}} e^{\beta\tau} dB(\tau)$$

are mutually independent. Then (applying this fact to subintervals of the intervals (t_j, t_{j+1}))

$$\int_{t_j}^{t_{j+1}} e^{-\beta\tau} d\tau \int_{t_j}^{t_j} e^{\beta\tau} dB(\tau), \quad j = 1, \dots, n$$

are mutually independent, and these repeated integrals are simply

$$B(t_{j+1}) - B(t_j) \quad j = 1, \dots, n-1.$$

The latter differences are therefore mutually independent, as was to be proved.

We shall need the following lemma.

LEMMA 3. Suppose that $a < 1$, and let x_0, x_1, \dots be mutually independent chance variables with a common distribution function. If there is a chance variable y with a Gaussian distribution such that the distribution function of $\sum_{i=0}^{n-1} a^{n-i} x_i$ approaches that of y as $n \rightarrow \infty$, then the x_j have Gaussian distributions.

Many of the hypotheses of the lemma are unnecessary, but its statement is general enough for our purposes, and the proof will apply to a situation to be discussed in §5, where the distribution of y will not be Gaussian. The hypotheses imply that the distribution of $\sum_{i=0}^n a^i x_i$ approaches that of $a^{n-1} y$ as $n \rightarrow \infty$. If $\varphi(t)$ is the characteristic function of x_1 and $\psi(t)$ that of y , writing $\sum_{i=1}^n a^i x_i$ in the form $ax_1 + \sum_{i=2}^n a^i x_i$ shows that

$$\psi(t) = \varphi(at) \cdot \psi(at).$$

Solving for φ we find that it is the characteristic function of a Gaussian distribution, as was to be proved.

In the physical picture under discussion, further conditions on the solution of (3.3) are known. In fact the Brownian movement is simply a visible example of molecular or near molecular movement. The general principles of such movements are therefore applicable, and the principle of equipartition of energy leads to the Maxwell distribution of velocities. Let k be the Boltzmann constant, and T the absolute temperature. We can formulate the significance of the Maxwell distribution (as much as we shall need it) as follows.

M_1 . *Tendency towards the Maxwell distribution.* Whatever the initial distribution of $u(0)$, the transition probabilities have the property that when $t \rightarrow \infty$ the distribution function of $u(t)$ converges to the Gaussian distribution function with mean 0 and variance kT/m . (Here m is the mass of the moving particle.)

M_2 . *Stability of the Maxwell distribution.* If $u(0)$ is independent of later residual random impacts, and if it has the Gaussian distribution described in M_1 , $u(t)$ will have this same distribution for every positive t .

These two statements are closely related, but neither apparently can be deduced from the other without further assumptions. Since these principles act the part of a *deus ex machina* in a discussion of the Langevin equation, we shall use them as little as possible. It will usually be sufficient to use a weakened form of M_1 :

M'_1 . There is an initial distribution of $u(0)$, such that the transition probabilities have the property that when $t \rightarrow \infty$ the distribution function of $u(t)$ converges to the Gaussian distribution function with mean 0 and variance kT/m . It is understood here as before that $u(0)$ is to be independent of later residual random impacts.

Conditions M_1 and M_2 restrict the possibilities for the $B(t)$ -process. In fact suppose that condition M'_1 is satisfied. Then (3.11) shows that

$$e^{-\beta t} \int_0^t e^{\beta \tau} dB(\tau)$$

is nearly Gaussian for large t , with mean 0 and variance kT/m . We write this integral as a sum, replacing t by nt :

$$(3.15) \quad e^{-\beta nt} \int_0^{nt} e^{\beta \tau} dB(\tau) = \sum_0^{n-1} e^{-\beta t(n-j)} x_j, \quad \wedge$$

where

$$(3.16) \quad x_j = \int_{jt}^{(j+1)t} e^{\beta(\tau-jt)} dB(\tau).$$

Since the $B(t)$ -process is a differential process, and is temporally homogeneous, the x_j are mutually independent, with identical distributions. According to the lemma, the right side of (3.15) cannot become Gaussian for large t unless the distribution of x_1 is Gaussian. Thus, since t is arbitrary in the above discussion,

$$\int_0^t e^{\beta \tau} dB(\tau)$$

has a Gaussian distribution for all s, t . Since the chance variables

$$(3.17) \quad \int_{jt/n}^{(j+1)t/n} e^{\beta\tau} dB(\tau), \quad j = 1, \dots, n$$

are mutually independent and Gaussian, the chance variable

$$(3.18) \quad \sum_{j=0}^{n-1} e^{-\beta jt/n} \int_{jt/n}^{(j+1)t/n} e^{\beta\tau} dB(\tau)$$

also has a Gaussian distribution. When n becomes infinite, (3.18) becomes $B(t) - B(0)$, with probability 1. The latter difference thus has a Gaussian distribution, with mean 0. The $B(t)$ -process therefore has finite second moments $\sigma^2(t) = t\sigma^2$ as defined in (3.4). According to (3.7) the last term in (3.11), which we now know has a Gaussian distribution, has mean 0 and variance $\sigma^2(1 - e^{-2\beta t})/2\beta$. Then $u(t) - e^{-\beta t}u(0)$ has this same distribution. The variance becomes $\sigma^2/2\beta$ when $t \rightarrow \infty$, and therefore, according to M'_1 , $\sigma^2 = 2\beta kT/m$. Thus condition M'_1 completely determines the $B(t)$ -process. We show next that condition M_2 determines this same $B(t)$ -process. In fact suppose condition M_2 is true, and assign to $u(0)$ the distribution of that condition. Then $u(0)$ is independent of the integral in (3.11), and in (3.11), $u(t)$ (which has a Gaussian distribution, according to condition M_2) is expressed as the sum of two independent chance variables, of which the first is Gaussian. The characteristic function of the second is the quotient of the characteristic functions of two Gaussian distributions, and is therefore the characteristic function of a Gaussian distribution. Thus the expression

$$(3.19) \quad e^{-\beta t} \int_0^t e^{\beta\tau} dB(\tau)$$

has a Gaussian distribution for all t , and this implies, as above, that $B(t) - B(0)$ has a Gaussian distribution, with variance $\sigma^2 t$. The variances on the right side of (3.11) add up to that on the left, giving an equation for σ^2 :

$$(3.20) \quad \frac{kT}{m} = e^{-2\beta t} \frac{kT}{m} + \frac{1 - e^{-2\beta t}}{2\beta} \sigma^2.$$

Then $\sigma^2 = 2\beta kT/m$ as above.

We can now finally derive the O. U. velocity process as the solution of the Langevin equation. Suppose the $B(t)$ -process is the one derived in the preceding paragraphs, and choose the chance variable $u(0)$ to be independent of the $B(t)$ -process for $t \geq 0$. Then $u(0)$ is independent of the integral in (3.11), and this means that the conditional distribution of $u(t)$ for $u(0) = u_0$ is Gaussian, with mean 0 and variance $kT(1 - e^{-2\beta t})/m$. Moreover, (3.11) implies

$$(3.21) \quad u(s+t) = u(s)e^{-\beta t} + e^{-\beta(s+t)} \int_s^{s+t} e^{\beta\tau} dB(\tau).$$

As we have seen, $u(s)$ is independent of the $B(t)$ -process as far as it appears in (3.21) and therefore is independent of the integral. Thus the transition

probabilities from s to $s + t$ are the same as those from 0 to t , which are precisely those of the O. U. process. Incidentally it follows that the full condition M_1 is satisfied. Finally, if $u(0)$ is not only supposed independent of the $B(t)$ -process, for $t \geq 0$, but also is supposed to have a Gaussian distribution with mean 0 and variance kT/m , the same will be true of $u(t)$ (as can be calculated from (3.11)) and condition M_2 is thus satisfied. We can summarize all our results as follows.

THEOREM 3. *Let the $B(t)$ -process be a temporally homogeneous differential process. Then (3.11) furnishes the solution of (3.3). The following conditions on the solution are equivalent.*

- (i) *The solution satisfies condition M'_1 .*
- (ii) *The solution satisfies condition M_1 .*
- (iii) *The solution satisfies condition M_2 .*
- (iv) *$B(t) - B(0)$ has a Gaussian distribution, with mean 0 and variance $\sigma^2 t = t2\beta kT/m$.*

If the above conditions are satisfied, $u(t) - e^{-\beta t}u(0)$ will have a Gaussian distribution with mean 0 and variance $kT(1 - e^{-2\beta t})/m$; if $u(0)$ is independent of the $B(t)$ -process for $t \geq 0$, $u(s)$ is independent of the $B(t)$ -process for $t \geq s$ for all $s > 0$, and the transition probabilities of the $u(t)$ -process are those of the O. U. velocity process. If in addition $u(0)$ has the Gaussian distribution with mean 0 and variance kT/m , the $u(t)$ -process becomes the O. U. process, with $m = 0$, $\sigma_0^2 = kT/m$.

The Langevin equation gives a physical interpretation to every property of the O. U. process. It is interesting to verify that as $h \rightarrow 0$ the correlation coefficient of the pair $B(s + h) - B(s)$, $u(t)$ (any s, t) goes to 0. In this sense then, $dB(s)$, the effect of the residual random impacts at time s , is independent of the velocity at any particular time t . Since in (3.11) $u(t)$ is written in terms of the $B(t)$ -process, $u(t)$ is of course not independent of this process.

We have written $u(t)$ in terms of the $B(t)$ -process. It is easy to write $x(t)$ in terms of the $B(t)$ process by combining (2.1) with (3.11):

$$(3.22) \quad x(t) = x(0) + \frac{1 - e^{-\beta t}}{\beta} u(0) + \frac{1}{\beta} \int_0^t [1 - e^{-\beta(t-\tau)}] dB(\tau).$$

Instead of finding the distributions of the displacement and velocity processes, and their correlations, as at the beginning of the paper, we could easily derive the desired results using (3.11) and (3.22). The various expectations can be calculated using (3.7).

In physical applications, the correlation function $E\{u(s)u(s + t)\}$ is sometimes wanted as a time average. Now the transformation S_h taking $B(t) - B(0)$ into $B(t + h) - B(h)$ preserves the $B(t)$ probability relations (temporal homogeneity), and the family of transformations $\{S_h\}$ is well known to be metrically transitive.⁹ Then applying the ergodic theorem to the function $u(0)u(h)$, considered as a function of the $B(t)$, we find that

⁹ Cf. for example Doob, (3) p. 125.

$$(3.23) \quad \lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t u(s)u(s+h) ds = E\{u(0)u(h)\} = \frac{kT}{m} e^{-\beta|h|},$$

with probability 1, that is for almost all functions $u(t)$. The ergodic theorem was applied to the $B(t)$ -process in essentially this way by Wiener ((18) p. 169) who has been interested in the harmonic analysis of functions like the $u(t)$ discussed here. The work of this paper verifies in this particular case the importance Wiener gave to the functions of the $B(t)$ -process of the type (3.5).

There is no difficulty in extending the above results to bound particles. For example, the Langevin equation of the harmonically bound particle is

$$(3.24) \quad \frac{du}{dt} = -\beta u - \omega^2 x + A(t),$$

which in our treatment becomes

$$(3.25) \quad du = -\beta u dt - \omega^2 x dt + dB.$$

The usual methods of solving the differential equation (3.24) are still applicable to (3.25) and again the distribution of u turns out to be Gaussian.¹⁰ The distribution of displacements is then obtained as above.

4. The $B(t)$ -impact process

When the $B(t)$ -process and the initial conditions on $u(0)$ are given, the solution $u(t)$ is determined by (3.11). Conversely if the solution $u(t)$ is known, $B(t)$ is determined by the equation

$$(4.1) \quad B(t) - B(0) = \beta \int_0^t u(s) ds + u(t) - u(0)$$

which is derived immediately from (3.3). The O. U. velocity distribution for the $u(t)$ -process can therefore be given only by the $B(t)$ -process described in §3. We shall investigate the possibility that a different choice of the $B(t)$ -process might have led to a different velocity process compatible with the known physical conditions like M_1 and M_2 . If we suppose that $u(0)$ can be chosen so that the velocity at each moment is independent of subsequent residual random impacts, then we have seen that the $B(t)$ -process must be differential, and is then uniquely determined by conditions M'_1 or M_2 . Any velocity process other than the O. U. process satisfying the Langevin equation and M'_1 or M_2 would therefore imply dependence between velocity and later residual impacts. This is really another way of saying that the frictional resistance cannot be considered as proportional to the velocity. Before going further we put a condition going back to Maxwell in its modern setting. We formulate a hypothesis M_3 as follows.

M_3 . In two or more dimensions (using any orthogonal axes) the velocity components are mutually independent.

¹⁰ Cf. Ornstein and Wijk (16) and Wijk (17). The $B(t, \Delta)$ used in these papers corresponds formally to our $dB(t)$. The difference is that it is possible to give a precise description of the $B(t)$ -distribution.

In conjunction with the following lemma, due to Kaç ((8) p. 278), hypothesis M_2 implies that all quantities linear in the displacement or velocity functions have Gaussian distributions.

LEMMA.¹¹ Let $(x_1, y_1), \dots, (x_n, y_n)$ be $2n$ chance variables with the property that the sets of chance variables

$$\{x_j \cos \theta + y_j \sin \theta, j = 1, \dots, n\} \{-x_j \sin \theta + y_j \cos \theta, j = 1, \dots, n\}$$

are mutually independent for each value of θ . Then (x_1, \dots, x_n) have an n -variate Gaussian distribution or a singular Gaussian distribution.

We can combine the Maxwell hypotheses to obtain another justification of the O. U. velocity process.

THEOREM 4. Let the $B(t)$ -process be any process such that the distribution of $B(t_2) - B(t_1)$ or of any quantity depending on such differences is unaffected by translations of the t -axis, and that the integral (3.5) can be defined as the limit in probability of the usual sums, with (3.6) valid. Then if $u(t)$ is defined by (3.11), and if conditions M_2 and M_3 are satisfied, the $B(t)$ -process must be precisely that finally obtained in §3, leading to the O. U. velocity process.

Suppose that condition M_2 is satisfied, and let $u(0)$ be fixed as in that condition. Just as in §3, (3.11) then implies that the integral

$$B^*(t) = \int_0^t e^{\beta\tau} dB(\tau)$$

has a Gaussian distribution with mean 0 and variance $(kT/m)(e^{2\beta t} - 1)$. If condition M_3 is also satisfied, $B^*(t_2) - B^*(t_1)$, and more generally any finite set of such differences, has a one or more dimensional Gaussian distribution. Using the fact that the distribution of $e^{-\beta s}[B^*(s+t) - B^*(s)]$ is the same as that of $B^*(t)$, in evaluating the expectations in the following equation

$$(4.2) \quad E\{B^*(s+t)^2\} = E\{[B^*(s) + [B^*(s+t) - B^*(s)]]^2\},$$

we find that $B^*(s) = B^*(s) - B^*(0)$ and $B^*(s+t) - B^*(s)$ are uncorrelated. These two variables are therefore independent. Going further, similar calculations show that any differences $B^*(t_2) - B^*(t_1)$, $B^*(s_2) - B^*(s_1)$ with $0 \leq s_1 < s_2 \leq t_1 < t_2$ are independent. Using the fact (derived from condition M_3) that any finite set of differences has a multivariate Gaussian distribution, the $B^*(t)$ -process is thus a differential process. This means, by a method we have used above, that the $B(t)$ -process is a differential process, leading to the O. U. velocity distribution, because condition M_2 is satisfied.

It is easily seen from counterexamples that Theorem 4 is no longer correct if condition M_1 is supposed instead of condition M_2 .

5. Velocity processes not subject to Maxwell's laws

In all the above work the role of the Maxwell velocity distribution has been fundamental. In certain studies, however, other distributions play a somewhat

¹¹ The result is stated slightly incorrectly by Kaç.

analogous role.¹² It is interesting to note that the Langevin equation can be solved to give a distribution whose transition probabilities are asymptotically any of the symmetric stable distributions classified by Lévy ((14) §30, §56, §57). Such a distribution has characteristic function

$$e^{-\sigma_0^2 |s|^\gamma}$$

where σ_0^2 is a positive parameter and $0 < \gamma \leq 2$. The Gaussian distribution is obtained when $\gamma = 2$. The parameter σ_0^2 plays the role of the variance, although the second moment is never finite when $\gamma < 2$. The velocity process we shall derive will be called the O. U. (γ) process. It is the O. U. process when $\gamma = 2$. The O. U. (γ) process can be described as follows.

1. The process is temporally homogeneous, that is translations of the t -axis do not affect the probability distributions.

2. The process is a Markoff process.

3. For each fixed t , $u(t)$ has a symmetric stable distribution with parameter value σ_0^2 , exponent γ . The conditional distribution of $u(s+t)$ for $u(s) = u_0$ is the stable distribution symmetric about $u_0 e^{-\beta t}$, with parameter value $\sigma_0^2(1 - e^{-\gamma\beta t})$ and exponent γ .

We can obtain this process as a solution of the Langevin equation by choosing the $B(t)$ -process properly. In fact, let the $B(t)$ -process be the temporally homogeneous differential process in which $B(s+t) - B(s)$ has a symmetric stable distribution with exponent γ and parameter value $\sigma^2 t$. Let $u(t)$ be the corresponding solution of the Langevin equation, given by (3.11). If y is the sum of two independent chance variables with stable symmetric distributions, having parameter values σ_1^2, σ_2^2 , and with the same exponent γ then y also has a symmetric stable distribution, with the same exponent, γ , and with parameter value $\sigma_1^2 + \sigma_2^2$. From this fact it is simple to check that the integral (3.5) in the present case has a symmetric stable distribution with exponent γ and parameter value.

$$\int_a^b |f(t)|^\gamma dt.$$

If $u(0)$ is given a symmetric stable distribution independent of the $B(t)$ -process for $t \geq 0$, with parameter value $\sigma^2/\gamma\beta$, the distribution of $u(t)$ can be calculated, using characteristic functions, and is found to be symmetric and stable, with exponent γ and parameter value $\sigma^2/\gamma\beta$. The $u(t)$ thus defined determines an O. U. (γ) process, with the above three properties, setting $\sigma_0^2 = \sigma^2/\gamma\beta$.

We shall not spend any time on the details of the analysis of the O. U. (γ) process, since the work runs parallel to that for the case $\gamma = 2$, already discussed. There are, however, a few essential differences. If $v(t)$ is determined by the equation

$$(1.2.1) \quad v(t) = t^{1/\gamma} u\left(\frac{1}{\gamma\beta} \log t\right), \quad t > 0,$$

¹² Cf. Holtzmark (7).

the $v(t)$ process can be analyzed using (3.11). The $v(t)$ -process has the same distribution as the $B(t)$ -process just described. The continuity properties of the velocity process can now be derived from those of the $v(t)$ -process, which are known. When $\gamma < 2$, the velocity function $u(t)$ is no longer a continuous function of t with probability 1, but is certain to have discontinuities. These discontinuities are however non-oscillatory (jumps).¹³ We omit the details of the analogue of Theorem 1.3. Theorem 2.1 is still true if $\gamma \geq 1$. The considerations of §3 have their obvious counterparts here. Lemma 3 played an essential role, but its statement and proof are correct if the variable y of the lemma is supposed to have a symmetric stable distribution and if the conclusion is that the x_j have a symmetric stable distribution with the same exponent as y .

UNIVERSITY OF ILLINOIS

AND

INSTITUTE FOR ADVANCED STUDY

BIBLIOGRAPHY

1. S. BERNSTEIN, Comptes Rendus de l'Académie des Sciences de l'URSS, N.S. (1934) pp. 1-9.
2. S. BERNSTEIN, Comptes Rendus de l'Académie des Sciences de l'URSS, N.S. (1934), pp. 361-364.
3. J. L. DOOB, Transactions of the American Mathematical Society 42 (1937), pp. 107-140.
4. J. L. DOOB, Duke Mathematical Journal 6 (1940), pp. 290-306.
5. J. L. DOOB, Transactions of the American Mathematical Society 47 (1940), pp. 455-486.
6. G. L. DE HAAS-LORENTZ, *Die Brownsche Bewegung und einige verwandte Erscheinungen*, Braunschweig (1913).
7. J. HOLTZMARK, Annalen der Physik 58 (1919), pp. 577-630.
8. M. KAÇ, American Journal of Mathematics 61 (1939), pp. 726-728.
9. A. KHINTCHINE, Mathematische Annalen 109 (1934), pp. 604-615.
10. A. KHINTCHINE, Ergebnisse der Mathematik 4 No. 3.
11. G. KRUTKOW, Physikalische Zeitschrift der Sowjet-Union 5 (1934), pp. 287-300.
12. P. LÉVY, Pisa Annali Series 2 vol. 3 (1934), pp. 337-366.
13. P. LÉVY, Pisa Annali Series 2 vol. 4 (1935), pp. 217-218.
14. P. LÉVY, *Théorie de l'addition des variables aléatoires*, Paris 1937.
15. L. S. ORNSTEIN AND G. E. UHLENBECK, Physical Review 36 (1930), pp. 823-841.
16. L. S. ORNSTEIN AND W. R. VAN WIJK, Physica 1 (1934), pp. 235-254, errata p. 966.
17. W. R. VAN WIJK, Physica 3 (1936), pp. 1111-1119.
18. N. WIENER AND R. E. A. C. PALEY, *Fourier Transforms in the Complex Domain*, American Mathematical Society Colloquium Publications Vol. XIX.

¹³ For further details, cf. Lévy (14) Chapter VII.

A NEW HOMOLOGY THEORY

By W. MAYER

(Received January 20, 1942)

In the classical homology theory one considers a sequence C^0, C^1, \dots of additive groups and homomorphisms $C^{i+1} \rightarrow C^i$ such that the induced homomorphisms $C^{i+2} \rightarrow C^i$ are trivial, i.e. C^{i+2} is mapped into the zero of C^i . In the present paper we propose to develop a new homology theory which also uses a sequence C^0, C^1, \dots of additive groups and homomorphisms $C^{i+1} \rightarrow C^i$. In this new theory, however, a fixed prime number p is chosen, and the induced homomorphisms $C^{i+p} \rightarrow C^i$ are the ones which are trivial. These groups for $p = 3$ were already considered at length, but by a different method, in the following papers: Mayer-Campbell, *Generalized Homology Groups*, Proc. Nat. Acad. Sci., U. S. A., 26, 655-656 (1940), and *Generalized Homology Groups*, to be published shortly in Revista de Matemáticas y Física Teórica.

1. Simplicial systems

We first define the new homology theory for a simplicial system i.e. a collection $\{\sigma\}$ of (non-oriented) simplexes such that the faces of any simplex $\sigma \in \Sigma$ also belongs to Σ . In addition to the simplexes of Σ , henceforth called *simple simplexes*, we shall consider simplexes with repeated vertices, or *generalized simplexes*

$$(1.1) \quad (P_1^{\alpha_1} P_2^{\alpha_2} \dots P_r^{\alpha_r})$$

where $P_i \neq P_k$ and the integers $\alpha_i \geq 1$ indicate the multiplicity of the corresponding vertex.

The simplex $(1, 1)$ shall belong to Σ if and only if the *simple simplex* $(P_1 P_2 \dots P_r)$ belongs to Σ . The dimension of the generalized simplex $(1, 1)$ is defined to be $\alpha_1 + \alpha_2 + \dots + \alpha_r - 1$.

Hereafter we use the symbol Σ to denote the extended system of all the simple and generalized simplexes thus obtained.

We now introduce, for a fixed prime number $p (\neq 1)$ the group C^n of n -chains, K^n , mod p . These n -chains K^n are made up of simplexes ρ^n of Σ of dimension n which may or may not be simple:

$$K^n = \sum_{j=1}^N a_j \rho_j^n.$$

A *boundary operator* F is first defined for simplexes of dimension > 0 by

$$(1.2) \quad F(P_1^{\alpha_1} \dots P_r^{\alpha_r}) = \sum_{i=1}^r \alpha_i (P_1^{\alpha_1} \dots P_{i-1}^{\alpha_{i-1}} P_{i+1}^{\alpha_{i+1}} \dots P_r^{\alpha_r}),$$

where vertices with exponent zero are crossed out, and the coefficients are reduced mod p . Just as in the classical theory, the formula

$$(1.3) \quad F(\sum a_j \rho_j^n) = \sum a_j F(\rho_j^n)$$

defines a homomorphism $C^n \rightarrow C^{n-1}$ for $n > 0$.

If the n -dimensional simplex (1.1) is written in the form

$$(1.4) \quad (P_1 P_2 \cdots P_{n+1})$$

where the vertices $P_1, P_2, \cdots, P_{n+1}$ need not be distinct, the formula (1.2) may be written

$$(1.5) \quad F(P_1 \cdots P_{n+1}) = \sum_{j=1}^{n+1} (P_1 \cdots P_{n+1})_j \quad n \geq 1,$$

where $(P_1 \cdots P_{n+1})_j$ is the face of $(P_1 \cdots P_{n+1})$ opposite the vertex P_j . Denote by $(P_1 \cdots P_{n+1})_{j_1 \cdots j_r}$ the face opposite the face $(P_{j_1} \cdots P_{j_r})$ and let $F^2(\) = F(F(\))$, \cdots , $F^i(\) = F(F^{i-1}(\))$. The formula (1.5) enables us to calculate rapidly the boundary of a boundary etc. Thus

$$F^2(P_1 \cdots P_{n+1}) = 2 \sum_{(i_1 i_2)} (P_1 \cdots P_{n+1})_{i_1 i_2}, \quad n \geq 2,$$

and in general,

$$(1.6) \quad F^i(P_1 \cdots P_{n+1}) = i! \sum_{(j_1 \cdots j_i)} (P_1 \cdots P_{n+1})_{j_1 \cdots j_i}, \quad n \geq i.$$

where the summation $(j_1 \cdots j_i)$ runs over all $\binom{n+1}{i}$ combinations of $1, 2, \cdots, n+1$. If $n \geq p$ then

$$(1.7) \quad F^p(P_1 \cdots P_{n+1}) = 0^{n-p},$$

where 0^r denotes the zero of C^r . If $i < p$ then in general $F^i(P_1 \cdots P_{n+1}) \neq 0^{n-i}$.

Thus, for $n \geq i$ the homomorphism F^i maps C^n into the zero of C^{n-i} if $i \geq p$.

2. q -cycles of dimension n

If $1 \leq q < \min \{p, n+1\}$, an n -chain K^n will be called an n -dimensional q -cycle (briefly a (q, n) -cycle) whenever

$$(2.1) \quad F^q(K^n) = 0^{n-q}.$$

By (1.3) the (q, n) -cycles form a subgroup Z_q^n of C^n .

For any $(n+p-q)$ -chain K^{n+p-q}

$$(2.2) \quad F^q[F^{p-q}(K^{n+p-q})] = F^p(K^{n+p-q}) = 0^{n-q}.$$

Hence the $(p-q)^{\text{th}}$ boundary of an $(n+p-q)$ -chain is always a (q, n) -cycle. These $(p-q)^{\text{th}}$ boundaries form a subgroup B_q^n of Z_q^n . The difference group

$$(2.3) \quad H_q^n = Z_q^n - B_q^n$$

will be called the q^{th} n -dimensional homology group of Σ (briefly the (q, n) -homology group). It is defined for $q \leq n$ since (2.1) has meaning only in this case. If $q > n$, so that (2.1) is not applicable, we define

$$(2.3) \quad Z_q^n = C^n, \quad q > n.$$

Thus every n -dimensional chain is a (q, n) -cycle when $q > n$. The group B_q^n of $(p - q)^{\text{th}}$ boundaries is defined for every n and every $q < p$ and is a subgroup of Z_q^n ; hence the group H_q^n is defined by (2.3) for every $n = 0, 1, 2, \dots$ and every q such that $1 \leq q < p$.

3. Regular and degenerate chains

We shall call a simplex σ *degenerate* if one or more of its vertices have a multiplicity greater than $p - 1$; otherwise it is said to be *regular*. A chain will be called *regular* if its simplexes are all regular and *degenerate* if its simplexes are all degenerate. Thus every chain K^n can be represented uniquely as the sum of a regular chain $K_{(r)}^n$ and a degenerate chain $K_{(d)}^n$

$$(3.1) \quad K^n = K_{(r)}^n + K_{(d)}^n.$$

Thereby we consider the zero chain as the only chain both regular and degenerate. From (1.2) and (1.3) it follows that the boundary of a regular chain is regular and the boundary of a degenerate chain is degenerate. Hence (for $n > 0$) (3.1) implies that

$$(3.2) \quad F(K^n) = F(K_{(r)}^n) + F(K_{(d)}^n),$$

where

$$(3.3) \quad [F(K^n)]_{(r)} = F(K_{(r)}^n), \quad [F(K^n)]_{(d)} = F(K_{(d)}^n).$$

Thus the simplicial system Σ is the "direct sum" of its two sub-systems $\Sigma_{(r)}$ (of regular chains) and $\Sigma_{(d)}$ (of degenerate chains):

$$\Sigma = \Sigma_{(r)} + \Sigma_{(d)}.$$

(3.4) THEOREM. The (q, n) -homology groups of Σ are isomorphic with the corresponding groups of $\Sigma_{(r)}$.

Let $K_{(r)}^n$ be a (q, n) -cycle of $\Sigma_{(r)}$; hence also a (q, n) -cycle of Σ . Let $\{K_{(r)}^n\}_{(r)}$ and $\{K_{(r)}^n\}$ denote its respective homology classes in $\Sigma_{(r)}$ and Σ . The mapping τ :

$$(3.5) \quad \{K_{(r)}^n\}_{(r)} \rightarrow \{K_{(r)}^n\}$$

defines a homomorphism $\bar{\tau}$ of the (q, n) homology group of $\Sigma_{(r)}$ into the (q, n) -homology group of Σ . The nucleus of $\bar{\tau}$ is the zero class of the (q, n) -homology group of $\Sigma_{(r)}$. In fact if $\{K_{(r)}^n\}_{(r)}$ belongs to the nucleus, then there is a chain K^{n+p-q} of Σ such that

$$(3.6) \quad F^{p-q}(K^{n+p-q}) = K_{(r)}^n.$$

If $K_{(r)}^{n+p-q}$ denotes the regular part of K^{n+p-q} , it follows then from (3.1), (3.2) and (3.3) that

$$(3.7) \quad F^{p-q}(K_{(r)}^{n+p-q}) = K_{(r)}^n.$$

Thus $\bar{\tau}$ is *univalent* (= an isomorphism with a subgroup). To prove (3.4) we merely need to show that $\bar{\tau}$ is a mapping "onto," i.e. that every class of the (q, n) -homology group of Σ contains a regular cycle. Let $\{K^n\}$ be such a class.

We may suppose that K^n is not regular, so that some vertex P appears in K^n with a multiplicity $> p - 1$. The chain K^n may then be written

$$(3.8) \quad K^n = K_1^n + K_2^n,$$

where

$$(3.9) \quad K_1^n = R^n + PR^{n-1} + \dots + P^{p-1}R^{n-p+1}, \quad K_2^n = P^p S^{n-p},$$

and the chains R^i do not have the vertex P . Since no $(n - q)$ -simplex can belong to the q -boundaries of both K_1^n and K_2^n it follows from $F^q(K^n) = 0^{n-q}$ that

$$(3.10) \quad F^q(K_1^n) = F^q(K_2^n) = 0^{n-q}.$$

The (q, n) -cycle K_2^n whose dimension n is greater than $p - 1$ lies in the closure of the star $\text{St } P$ of P . Hence (Appendix I) K_2^n belongs to the zero class of the (q, n) -homology group of Σ and hence, by (3.8), K_1^n belongs to the homology class of K^n . Thus the removal of those simplexes of a cycle K^n which contain a vertex P with a multiplicity greater than $p - 1$ does not alter the homology class of K^n . Obviously $K_{(r)}^n$ is obtained from K^n by repetition of this process. Hence the regular (q, n) -cycle $K_{(r)}^n$ belongs to the homology class of K^n . This completes the proof of (3.4).

4. Invariance under Subdivision

We shall consider here the effect of a certain elementary subdivision of the simplicial system Σ with respect to the new homology groups. Let (ab) be a 1-simplex of the not-extended system Σ and γ the "midpoint" of (ab) . We form a new simplicial system, $\Sigma(ab)$, which has all the simple simplexes of Σ except those which contain both of the vertices a and b . A simplex $\Delta(ab)$ where Δ is a simplex free from a and b and may be vacuous, is replaced by the simplexes

$$(4.1) \quad \Delta(a\gamma), \quad \Delta(b\gamma), \quad \Delta\gamma,$$

where the first two have the same dimension as $\Delta(ab)$ and the last has a dimension lower by one. The simple simplexes of Σ which do not contain the face (ab) the simplexes (4.1) and their generalized simplexes, constitute the simplicial system $\Sigma(ab)$.

In the construction of $\Sigma(ab)$ from Σ we limited ourselves to simple simplexes because every simplicial system is determined by its simple simplexes.

(4.2) **THEOREM:** *The simplicial systems Σ and $\Sigma(ab)$ have isomorphic (q, n) -homology groups.*

First we construct a subsystem Σ^* of $\Sigma(ab)$ which is isomorphic to Σ . The n -chains $*C^n$ of Σ^* are generated by n -chains of $\Sigma(ab)$ of the form

$$(4.3) \quad \Delta(a'\gamma^\mu - \gamma'^{+\mu} + \gamma'b^\mu),$$

where $\nu, \mu = 0, 1, 2, \dots$ are non-negative integers and Δ are simplexes of $\Sigma(ab)$ free from the vertices a, b and γ and may be vacuous. It is easy to verify that $F(*C^{n+1}) \subset *C^n$ such that Σ^* is indeed a subsystem (not with a simplicial basis) of $\Sigma(ab)$. Furthermore, by the 1:1 correspondence

$$(4.4) \quad \Delta(a'\gamma^\mu - \gamma'^{+\mu} + \gamma'b^\mu) \leftrightarrow \Delta(a'b^\mu)$$

between the bases of Σ^* and Σ we establish isomorphisms between the groups of the n -chains $*C^n$ and $C^n(\Sigma)$ for $n = 0, 1, \dots$, which for $n = 1, 2, \dots$ commute with the boundary operator F , thus showing the isomorphism of the systems Σ and Σ^* . We prove this last statement for the basic-chains (4.4).

It is trivial when $\nu = 0$ or $\mu = 0$. We therefore assume that both ν and μ are ≥ 1 . Then (from the rule of Appendix I for taking boundaries of product chains)

$$(4.5) \quad \left\{ \begin{aligned} F[\Delta(a'b^\mu)] &= (a'b^\mu)F(\Delta) + \Delta[\nu a'^{-1}b^\mu + \mu a'b^{\mu-1}] \\ &\leftrightarrow (a'\gamma^\mu - \gamma'^{+\mu} + \gamma'b^\mu)F(\Delta) + \Delta[\nu(a'^{-1}\gamma^\mu + \gamma'^{+\mu-1} + \gamma'^{-1}b^\mu) \\ &\quad + \mu(a'\gamma^{\mu-1} - \gamma'^{+\mu-1} + \gamma'b^{\mu-1})] = F[\Delta(a'\gamma^\mu - \gamma'^{+\mu} + \gamma'b^\mu)]. \end{aligned} \right.$$

Hence Σ and Σ^* are isomorphic.

Now we define a homomorphism of $C^n(\Sigma(ab))$ into $*C^n$ by the mapping

$$(4.6) \quad \left\{ \begin{aligned} \Delta a'\gamma^\mu &\rightarrow \Delta a'^{+\mu} \\ \Delta \gamma'b^\mu &\rightarrow \Delta(a'\gamma^\mu - \gamma'^{+\mu} + \gamma'b^\mu), \end{aligned} \right.$$

of the basic chains (simplexes) of $C^n(\Sigma(ab))$ into basic chains of $*C^n$. As before, Δ denotes simplexes free from the vertices a, b and γ and ν and μ are \geq zero. This homomorphism also commutes with the boundary operator F . In fact

$$(4.7) \quad \left\{ \begin{aligned} F[\Delta a'\gamma^\mu] &= a'\gamma^\mu F(\Delta) + \Delta[\nu a'^{-1}\gamma^\mu + \mu a'\gamma^{\mu-1}] \\ &\rightarrow a'^{+\mu}F(\Delta) + \Delta[(\nu + \mu)a'^{+\mu-1}] = F(\Delta a'^{+\mu}), \end{aligned} \right.$$

and

$$(4.8) \quad \left\{ \begin{aligned} F[\Delta \gamma'b^\mu] &= \gamma'b^\mu F(\Delta) + \Delta[\nu \gamma'^{-1}b^\mu + \mu \gamma'b^{\mu-1}] \\ &\rightarrow (a'\gamma^\mu - \gamma'^{+\mu} + \gamma'b^\mu)F(\Delta) + \Delta[\nu(a'^{-1}\gamma^\mu - \gamma'^{+\mu-1} + \gamma'^{-1}b^\mu) \\ &\quad + \mu(a'\gamma^{\mu-1} - \gamma'^{+\mu-1} + \gamma'b^{\mu-1})] = F[\Delta(a'\gamma^\mu - \gamma'^{+\mu} + \gamma'b^\mu)]. \end{aligned} \right.$$

The homomorphism (4.6) therefore maps (q, n) -cycles into (q, n) -cycles and $(p - q)^{\text{th}}$ boundaries into $(p - q)^{\text{th}}$ boundaries. In particular it determines a

homomorphism of the (q, n) -homology groups which we now show to be an isomorphism.

The homomorphism (4.6) leaves the chains of $*C^n$ unaltered. We show this also for the basic chains (4.3) of $*C^n$. In fact we have here

$$\Delta(a^r\gamma^\mu - \gamma^{r+\mu} + \gamma^rb^\mu) \rightarrow \Delta[a^{r+\mu} - a^{r+\mu} + a^r\gamma^\mu - \gamma^{r+\mu} + \gamma^rb^\mu].$$

Now let K^n be a (q, n) -cycle of $*C^n$ and let $\{K^n\}^*$ and $\{K^n\}$ denote its homology classes in $*H_q^n$ and $H_q^n(\Sigma(ab))$ respectively. Under the homomorphism of the homology groups defined by (4.6) the image of $\{K^n\}$ is $\{K^n\}^*$. Thus the image of this homomorphism is the (q, n) -homology group $*H_q^n$ of Σ^* itself.

Suppose now that K^n is a (q, n) -cycle of $C^n(\Sigma(ab))$ which is mapped by the homomorphism (4.6) into the zero class of $*H_q^n$, i.e. the image cycle $*K^n$ of K^n is a $(p - q)^{\text{th}}$ boundary of a chain of $*C^{n+p-q}$. By (4.6) the (q, n) -cycle $K^n - *K^n$ is a linear combination of basis chains of the form

$$(4.9) \quad \Delta(a^r\gamma^\mu - a^{r+\mu}), \quad \Delta(a^r\gamma^\mu - \gamma^{r+\mu}).$$

The sum of the coefficients of $K^n - *K^n$ is therefore zero.

Furthermore, since the chains (4.9) lie in the closure of $\text{St } a$ of $\Sigma(ab)$, it follows [Appendix I] that the cycle $K^n - *K^n$ lies in the zero class of its homology group, i.e. is a $(p - q)^{\text{th}}$ boundary. Since $*K^n$ is a $(p - q)^{\text{th}}$ boundary, it follows that K^n itself is a $(p - q)^{\text{th}}$ boundary of a chain of $C^n(\Sigma(ab))$.

Thus the homomorphism of $H_q^n(\Sigma(ab))$ onto $*H_q^n$ is univalent and therefore an isomorphism. Hence, referring to the isomorphism (4.4), the simplicial systems Σ and $\Sigma(ab)$ have isomorphic (q, n) -homology groups. This isomorphism, which is a combination of the isomorphisms (4.6) and (4.4), is obviously generated by the simplicial mapping of $\Sigma(ab)$ into Σ in which γ is mapped into a and the other vertices remain unaltered. Collecting the results, we have:

The simplicial mapping of $\Sigma(ab)$ into Σ , in which γ is mapped into a (or b) and the other vertices remain unaltered, generates an isomorphism of the (q, n) -homology groups.

5. The new homology groups for topological spaces

The passage from finite simplicial systems to arbitrary topological spaces is similar to the procedure utilized by Čech for ordinary homology groups. We shall suppose the reader familiar with that method, and for details refer him to Čech's initial paper: *Théorie générale de l'homologie dans un espace quelconque*, Fundam. Mat. 19 (1932), 149–183, or to the full exposition of the theory in the forthcoming book by Lefschetz, *Algebraic Topology*, in the Colloquium Series, Chap. VII, §1. This book will be referred to in the sequel as “L,” and its general terminology will be utilized in the present section. In particular, here also a topological space designates a space which satisfies all but the separation axioms for Hausdorff spaces (L, Ch. I, No. 6).

The general argument in the Čech theory runs as follows: Let $\{U_\lambda\}$ be the finite open coverings of a topological space \mathfrak{R} , Φ_λ the nerve of U_λ , and if U_λ

refines U_μ let π_μ^λ be a projection by inclusion $\Phi_\lambda \rightarrow \Phi_\mu$ (see L, Ch. VII, 1.3). The projections π_μ^λ for given λ, μ , induce a unique simultaneous homomorphism π_μ^λ of the homology groups H_λ^n of Φ_λ into the corresponding groups of Φ_μ thus giving rise to inverse systems: $S^n = \{H_\lambda^n; \pi_\mu^\lambda\}$, and the corresponding groups of \mathfrak{R} are defined as $H^n = \lim S^n$.

The same argument may be repeated for the new homology groups. The only step which is new is the explicit proof that π_μ^λ is unique. As in (L, Ch. VII, 1.4) any two projections $\pi_\mu^\lambda, \pi'_\mu^\lambda$ are shown to be "prismatically related" (in the sense of L, Ch. IV, 16.2). The explicit deduction of the uniqueness of π_μ^λ from this property is given in Appendix II. As a consequence we shall have here also the inverse system $S_q^n = \{H_q^n(\phi_\lambda); \pi_\mu^\lambda\}$ and define for \mathfrak{R} : $H_q^n = \lim S_q^n$.

6. Application to finite polyhedra

Let $|\Sigma|$ be a finite simplicial polyhedron, where Σ is a finite Euclidean complex in the sense of (L, Ch. III, 6.9). We have thus the topological groups $H_q^n(|\Sigma|)$ of the space $|\Sigma|$ in the sense just defined, and also the combinatorial groups $H_q^n(\Sigma)$ of the simplicial system Σ . The proof of the isomorphism of the two is carried out as in (L, Ch. VIII, 10). The only modification made is the following: Instead of taking the successive barycentric subdivisions we choose successive subdivisions $\Sigma, \Sigma^1, \Sigma^2, \dots, \Sigma^n, \dots$ where Σ^{n+1} is deduced from Σ^n by introducing the "midpoint" γ of one of the largest of its one-simplexes (ab) .

In the notation of §4 we have $\Sigma^{n+1} = \Sigma^n(ab)$. We have seen in §4 that for the simplicial projection

$$\pi_n^{n+1}: \Sigma^{n+1} \rightarrow \Sigma^n,$$

which leaves unaltered all vertices of Σ^{n+1} but γ , which vertex is mapped in one of the vertices a or b , isomorphisms between the new homology groups result. To complete the parallel with the treatment *loc. cit.* it suffices to observe that the maximal diameter of Σ^n has a length converging to zero for $n \rightarrow \infty$.

This follows from the fact that in constructing Σ^{n+1} from Σ^n we choose the midpoint of one of the largest of the one-simplexes of Σ^n .

We have then the analogue of (L, Ch. VIII, 10.1):

(6.1) THEOREM. *The groups $H_q^n(\Sigma)$ of the simplicial system Σ are isomorphic with the corresponding groups of the polyhedron $|\Sigma|$. Therefore the $H_q^n(\Sigma)$ are topological invariants of $|\Sigma|$. That is to say, if two polyhedra $|\Sigma|, |\Sigma_1|$ are homeomorphic, the corresponding (combinatorial) groups H_q^n are the same.*

7. Appendix I

Let K^n be an n -dimensional chain of the simplicial system Σ , P a vertex of Σ and $\text{St } P$, the star of P in Σ .

(7.1) THEOREM: *If $n \neq q - 1$ then every (q, n) -cycle K^n which lies in the closure of $\text{St } P$ is a $(p - q)^{\text{th}}$ boundary; if $n = q - 1$ then a (q, n) -cycle K^n which lies in the closure of $\text{St } P$ is a $(p - q)^{\text{th}}$ boundary if and only if the sum of its coefficients is zero (mod p).*

Before proving this theorem we define the product-chain of the chains

$$K_1^\nu = \sum_{\lambda} a_{1\lambda} \rho_{1\lambda}^\nu, \quad K_2^\mu = \sum_{\tau} a_{2\tau} \rho_{2\tau}^\mu$$

by

$$(7.2) \quad K_1^\nu K_2^\mu = \sum_{\lambda, \tau} a_{1\lambda} a_{2\tau} \rho_{1\lambda}^\nu \rho_{2\tau}^\mu,$$

where $\rho_{1\lambda}^\nu \rho_{2\tau}^\mu$ is the $(\nu + \mu + 1)$ -dimensional simplex containing the vertices of both simplexes $\rho_{1\lambda}^\nu$ and $\rho_{2\tau}^\mu$, if and only if the product chain (7.2) belongs to Σ . If μ and ν are positive integers, then $F(K_1^\nu)$ and $F(K_2^\mu)$ are defined and

$$(7.3) \quad F(K_1^\nu K_2^\mu) = K_2^\mu F(K_1^\nu) + K_1^\nu F(K_2^\mu).$$

We can extend this rule to cover the cases of $\nu \leq 0$ or $\mu \leq 0$ (i.e. chains of zero or "negative dimensions")¹ by defining

$$(7.4) \quad F(P^0) = 1, \quad P^0 \text{ a zero-dimensional simplex,}$$

and

$$(7.5) \quad F(1) = 0, \quad F(0) = 0.$$

REMARK: If $q > n$ then $F^q(K^n) = 0$ no longer characterizes the (q, n) -cycles (cf. §2). Because then every n -dimensional chain is a q -cycle whereas $F^q(K^n) = 0$ is not always true if $q = n + 1$.

We now return to the proof of (7.1). If K^n is an n -dimensional chain which lies in the closure of $\text{St } P$, we define for it the (linear) operators D_ν , $\nu = 1, 2, \dots, p$, by the formula

$$(7.6) \quad D_\nu(K^n) = (-1)^{\nu-1}(\nu-1)!P^{p-\nu}K^n, \quad \nu = 1, \dots, p.$$

In this formula $0!$ and P^0 are to be replaced, as usual, by 1. By (7.3) and (7.6) we have, for $\nu = 1, 2, \dots, p-1$

$$(7.7) \quad F D_\nu(K^n) = (-1)^\nu \nu! P^{p-\nu-1} K^n + (-1)^{\nu-1}(\nu-1)! P^{p-\nu} F(K^n).$$

Hence

$$(7.8) \quad (F D_\nu - D_\nu F) K^n = D_{\nu+1} K^n, \quad \nu = 1, \dots, p-1.$$

Thus, in terms of operators

$$(7.9) \quad F D_\nu - D_\nu F = D_{\nu+1}, \quad \nu = 1, \dots, p-1.$$

¹ In this section and the next we shall make use for formal reasons of chains of negative dimensions.

To the chain-groups C^n , $n = 0, 1, 2, \dots$ of §2 we add the chain-groups C^{-n} , $n = 1, 2, \dots$ of chains of negative dimensions. C^{-1} (by definition) is the group of the rest-classes modulo p and the groups C^{-n} , $n = 2, 3, \dots$ consist of the zero-element only.

(7.4) and (7.5) define the homomorphisms $C^0 \rightarrow C^{-1}$, $C^{-1} \rightarrow C^{-2}$, \dots . The relation (7.3) then is valid for any ν and $\mu = 0, \pm 1, \pm 2, \dots$.

By eliminating D_2, D_3, \dots, D_r from the first r of the formulae (7.9), we find

$$(7.10) \quad D_{r+1} = \sum_{j=0}^r (-1)^j \binom{r}{j} F^{r-j} D_1 F^j.$$

Thus, for $r = 1, 2, \dots, p-1$ we have

$$(7.11) \quad \left\{ \sum_{j=0}^r (-1)^j \binom{r}{j} F^{r-j} D_1 F^j \right\} K^n = (-1)^r r! P^{p-r-1} K^n.$$

In particular, when $r = p-1$,

$$(7.12) \quad \left\{ \sum_{j=0}^{p-1} (-1)^j \binom{p-1}{j} F^{p-1-j} D_1 F^j \right\} K^n = (p-1)! K^n.$$

But p is a prime number so that $(p-1)! \equiv -1$ and $\binom{p-1}{j} \equiv (-1)^j \pmod{p}$, $j = 0, 1, \dots, p-1$. Hence, for every chain K^n in the closure of $\text{St } P$,

$$(7.13) \quad \begin{cases} F^{p-1} D_1 K^n + F^{p-2} D_1 F K^n + \dots + F D_1 F^{p-2} K^n + D_1 F^{p-1} K^n \\ = -K^n. \end{cases}$$

Now let K^n be a (q, n) -cycle in the closure of $\text{St } P$. If $q \leq n$ or $q > n+1$ then $F^q(K^n) = 0$ by (2.1), (7.4) and (7.5). Hence, by (7.13), if $q \neq n+1$

$$(7.14) \quad \begin{aligned} -K^n &= F^{p-1} D_1 K^n + \dots + F^{p-q} D_1 F^{q-1} K^n \\ &= F^{p-q} (F^{q-1} D_1 K^n + \dots + D_1 F^{q-1} K^n). \end{aligned}$$

This proves the first assertion of (7.1).

If $q = n+1$ then by (7.4), $F^q(K^n) = F^{n+1}(K^n)$ is the sum of the coefficients of $F^n(K^n)$. Suppose that

$$(7.15) \quad K^n = \sum a_\alpha (P_{\alpha_1} \dots P_{\alpha_{n+1}}).$$

Then, by (1.6),

$$(7.16) \quad F^n(K^n) = n! \sum a_\alpha \sum_{(j_1 \dots j_n)} (P_{\alpha_1} \dots P_{\alpha_{n+1}})_{j_1 \dots j_n}.$$

So that the sum of the coefficients of $F^n(K^n)$ is $(n+1)! \sum a_\alpha$. But $n+1 = q < p$, hence $F^{n+1}(K^n)$ is zero if and only if $\sum a_\alpha \equiv 0 \pmod{p}$. If this is the case, then (7.14) holds and K^n is a $(p-q)$ th boundary. On the other hand if K^n is a $[p - (n+1)]$ th boundary

$$(7.17) \quad K^n = F^{p-n-1}(K_1^{p-1}),$$

and if

$$(7.18) \quad K_1^{p-1} = \sum a_\alpha (P_{\alpha_1} \dots P_{\alpha_p})$$

then, (1.6),

$$(7.19) \quad K^n = (p-n-1)! \sum a_\alpha \sum_{(j_1 \dots j_{p-n-1})} (P_{\alpha_1} \dots P_{\alpha_p})_{j_1 \dots j_{p-n-1}}$$

has as the sum of coefficients

$$(7.20) \quad (p - n - 1)! \sum a_\alpha \binom{p}{p - n - 1}$$

and this sum is zero modulo p . This completes the proof of (7.1).

8. Appendix II

Let $\tilde{\Sigma}$ and Σ be simplicial systems and let f and g be simplicial mappings of $\tilde{\Sigma}$ into Σ . Let \tilde{A}_ν , $\nu = 1, 2, \dots, N$, denote the vertices of $\tilde{\Sigma}$ and let $A_\nu = f(\tilde{A}_\nu)$ and $A'_\nu = g(\tilde{A}_\nu)$.

(8.1) THEOREM: *If, for every simplex $(\tilde{A}_1 \cdots \tilde{A}_n)$ of $\tilde{\Sigma}$ the simplex $(A_1 \cdots A_n A'_1 \cdots A'_n)$ belongs to Σ , then f and g determine the same homomorphism of $H_q^n(\tilde{\Sigma})$ into $H_q^n(\Sigma)$.*

In proving this theorem we first order the vertices of $\tilde{\Sigma}$ so that the vertices of every simple simplex receive a certain definite ordering. If the simplex $(\tilde{A}_1 \cdots \tilde{A}_n)$ is not simple, then $i < k$ implies either that $\tilde{A}_i = \tilde{A}_k$ or that \tilde{A}_i precedes \tilde{A}_k in the given ordering of the vertices of $\tilde{\Sigma}$.

We now define p linear operators D_0, D_1, \dots, D_{p-1} , each one of which maps chains of $\tilde{\Sigma}$ into chains of Σ . It is sufficient to define these operators for the simplexes of $\tilde{\Sigma}$ and this is done by the formulae

$$(8.2) \quad \left\{ \begin{aligned} D_r(\tilde{A}_1 \cdots \tilde{A}_n) &= (-1)^r r! \left[\sum_{i=1}^n (A_1 \cdots A_{i-1} A_i A'_i{}^{p-r-1} A'_{i+1} \cdots A'_n) \right. \\ &\quad \left. - \sum_{i=1}^n (A_1 \cdots A_{i-1} A_i{}^{p-r} A'_{i+1} \cdots A'_n) \right]. \end{aligned} \right.$$

Since we have to apply these operators also to boundaries of chains, we have to extend the above definition to chains K of "negative dimensions," in which case $D_r(K)$ shall be zero.

If $r < p - 1$ we have

$$(8.3) \quad \left\{ \begin{aligned} FD_r(\tilde{A}_1 \cdots \tilde{A}_n) &= (-1)^{r+1} (r+1)! \left[\sum_{i=1}^n (A_1 \cdots A_i A'_i{}^{p-r-2} \cdots A'_n) \right. \\ &\quad \left. - \sum_{i=1}^n (A_1 \cdots A_{i-1} A'_i{}^{p-r-1} \cdots A'_n) \right] \\ &\quad + (-1)^r r! \left[\sum_{i=1}^n \sum_{j \neq i} (A_1 \cdots A_i A'_i{}^{p-r-1} \cdots A'_n)_j \right. \\ &\quad \left. - \sum_{i=1}^n \sum_{j \neq i} (A_1 \cdots A_{i-1} A'_i{}^{p-r} \cdots A'_n)_j \right]. \end{aligned} \right.$$

Furthermore

$$(8.4) \quad \sum_{j=1}^n D_r(\tilde{A}_1 \cdots \tilde{A}_n)_j = D_r \sum_{j=1}^n (\tilde{A}_1 \cdots \tilde{A}_n)_j = D_r F(\tilde{A}_1 \cdots \tilde{A}_n).$$

Hence from (8.2), (8.4)

$$(8.5) \quad \begin{aligned} D_r F(\tilde{A}_1 \cdots \tilde{A}_n) &= (-1)^r r! \left[\sum_{j=1}^n \sum_{i \neq j} (A_1 \cdots A_i A'_i{}^{p-r-1} \cdots A'_n)_i \right. \\ &\quad \left. - \sum_{j=1}^n \sum_{i \neq j} (A_1 \cdots A_{i-1} A'_i{}^{p-r} \cdots A'_n)_i \right]. \end{aligned}$$

This combined with (8.3) gives

$$(8.6) \quad (FD_r - D_r F)(\tilde{A}_1 \cdots \tilde{A}_n) = D_{r+1}(\tilde{A}_1 \cdots \tilde{A}_n), \quad r < p-1,$$

which formula also holds for chains of dimension ≤ 0 . Hence, just as in Appendix I,

$$(8.7) \quad D_r = \sum_{j=0}^r (-1)^j \binom{r}{j} F^{p-j} D_0 F^j, \quad r < p.$$

In particular, when $r = p-1$ (since $(p-1)! \equiv -1$, $\binom{p-1}{j} \equiv (-1)^j \pmod{p}$) we arrive at

$$(8.8) \quad \begin{aligned} \left(\sum_{j=0}^{p-1} F^{p-j-1} D_0 F^j \right) (\tilde{A}_1 \cdots \tilde{A}_n) &= - \sum_{i=1}^n (A_1 \cdots A_i A'_{i+1} \cdots A'_n) \\ &+ \sum_{i=1}^n (A_1 \cdots A_{i-1} A'_i \cdots A'_n) = (A'_1 \cdots A'_n) - (A_1 \cdots A_n). \end{aligned}$$

Hence for any $(n-1)$ -dimensional chain \tilde{K}^{n-1} of $\tilde{\Sigma}$ we have

$$(8.9) \quad \left(\sum_{j=0}^{p-1} F^{p-j-1} D_0 F^j \right) \tilde{K}^{n-1} = K'^{n-1} - K^{n-1},$$

where $K^{n-1} = f(\tilde{K}^{n-1})$ and $K'^{n-1} = g(\tilde{K}^{n-1})$.

Now let \tilde{K}^{n-1} be a q -cycle, then

$$(8.10) \quad D_0 F^q \tilde{K}^{n-1}, \dots, D_0 F^{p-1} \tilde{K}^{n-1}$$

are all zero since either the $F^q \tilde{K}^{n-1}, \dots, F^{p-1} \tilde{K}^{n-1}$, are zero or they are of "negative dimension."

Hence

$$(8.11) \quad K'^{n-1} - K^{n-1} = F^{p-q} \left[\left(\sum_{j=0}^{q-1} F^{q-j-1} D_0 F^j \right) \tilde{K}^{n-1} \right],$$

thus $K'^{n-1} - K^{n-1}$ is a $(p-q)$ -boundary.

Hence K'^{n-1} and K^{n-1} determine the same element of $H_q^{n-1}(\Sigma)$. This completes the proof of (8.1).

REMARK: The above operations D_r generalize in obvious manner the so-called "homotopy-operator" D of (L, Ch. IV, No. 14).

ON THE DIFFERENTIAL EQUATIONS OF THE SIMPLEST BOUNDARY-LAYER PROBLEMS

BY HERMANN WEYL

(Received February 10, 1942)

1. The central boundary-value problem and its hydrodynamic interpretation

In the theory of viscous fluids the following non-linear boundary-value problem for a function $w(z)$ of a real variable $z \geq 0$ involving two constants $k > 0$ and $\lambda \geq 0$ plays an important part:

$$(A_\lambda) \quad \begin{cases} w''' + 2ww'' + 2\lambda(k^2 - w^2) = 0 & \text{for } z \geq 0; \\ w(0) = w'(0) = 0, \quad w'(z) \rightarrow k & \text{for } z \rightarrow \infty. \end{cases}$$

We consider λ as a given constant, but k as a variable parameter. A mathematically satisfactory proof of its solvability has never been given, although various numerical devices, including V. Bush's differential analyzer, have been set at work on it. We shall here give a complete solution of the problem,¹ first for the two special values $\lambda = 0$ and $\lambda = \frac{1}{2}$ by a process of alternating approximations, rapidly converging and thus well suited for numerical computations (§§2, 4, 5), and then approach the general case (§§6, 7) by the method of fixed points of transformations in a functional space,—which is considerably less amenable to calculation. In between (§3) the first method will be applied to certain boundary-value problems closely related to (A_0) .

There are available two hydrodynamic interpretations of (A_λ) . Consider first the steady flow of an incompressible viscous fluid of constant density ρ and kinematic viscosity ϵ^2 filling the half $z > 0$ of an m -dimensional Euclidean space with the Cartesian coordinates x_1, \dots, x_m and the cylindrical coordinates

$$r = (x_1^2 + \dots + x_{m-1}^2)^{\frac{1}{2}}, \quad z = x_m.$$

If cylindrical symmetry prevails and hence the radial (r) and vertical (z) components u, v of velocity as well as the pressure p depend on r, z only, then the following differential equations obtain for $z > 0$:

$$(1) \quad \begin{cases} \frac{\partial u}{\partial r} + \frac{m-2}{r}u + \frac{\partial v}{\partial z} = 0; \\ u \frac{\partial u}{\partial r} + v \frac{\partial u}{\partial z} + \frac{\partial p}{\partial r} = \epsilon^2 \left(\Delta u - \frac{m-2}{r^2}u \right), \\ u \frac{\partial v}{\partial r} + v \frac{\partial v}{\partial z} + \frac{\partial p}{\partial z} = \epsilon^2 \Delta v \end{cases}$$

¹ See the author's preliminary notes in Proc. Nat. Acad. Sci. 27, 1941, pp. 578-583, and 28, 1942, pp. 100-102.

where the Laplace operator Δ is defined by

$$\Delta\varphi = \frac{\partial^2 \varphi}{\partial r^2} + \frac{m-2}{r} \frac{\partial \varphi}{\partial r} + \frac{\partial^2 \varphi}{\partial z^2}.$$

These equations are to be combined with the boundary conditions

$$u \rightarrow 0, \quad v \rightarrow 0 \quad \text{for} \quad z \rightarrow 0.$$

For an ideal fluid, $\epsilon = 0$, we have this simple solution:

$$u_0(r, z) = \frac{2k}{m-1} r, \quad v_0(r, z) = -2kz,$$

$$p_0 = \text{const.} - \frac{1}{2}(u_0^2 + v_0^2) = \text{const.} - 2k^2 \left\{ \frac{r^2}{(m-1)^2} + z^2 \right\}$$

arising from the harmonic velocity potential

$$\varphi = k \left(\frac{x_1^2 + \cdots + x_{m-1}^2}{m-1} - x_m^2 \right) = k \left(\frac{r^2}{m-1} - z^2 \right)$$

and involving an arbitrary positive constant k . As is necessary, the vertical (though not the radial) component velocity vanishes along the boundary $z = 0$. The Navier-Stokes equations (1) for the viscous fluid possess a solution of the form

$$u = \frac{2}{m-1} r F''(z), \quad v = -2F(z), \quad p = \text{const.} - 2k^2 \left\{ \left(\frac{r}{m-1} \right)^2 + L(z) \right\}$$

which approaches the solution u_0, v_0, p_0 for $z \rightarrow \infty$. The first equation (1) is identically satisfied, the second and third yield

$$\epsilon^2 F'''' + 2FF'' + \frac{2}{m-1} (k^2 - F'^2) = 0$$

and

$$(2) \quad k^2 L' = 2FF' + \epsilon^2 F''$$

respectively. Setting $F(\epsilon z) = \epsilon \cdot w(z)$ we obtain the equations (A_λ) with $\lambda = 1/(m-1)$ from which the viscosity constant has disappeared, so that w is independent of ϵ . Equation (2) in integrated form gives

$$k^2 \cdot L(\epsilon z) = \epsilon^2 \{w'(z) + w^2(z)\}.$$

Our solution describes approximately the flow of a viscous fluid around an obstacle with a blunt nose in the neighborhood of the forward stagnation point. The cases of physical interest, $m = 2$ and 3 , i.e. $\lambda = 1$ and $\frac{1}{2}$, have been treated by Hiemenz and Homann respectively.²

² Hiemenz, Döngler's Polytech. Jour. 326, 1911, pp. 321-326. F. Homann, Zeitschr. angew. Math. Mech. 16, 1936, p. 153.

Let the subscript ϵ in $u_\epsilon, v_\epsilon, p_\epsilon$ indicate dependence of our flow on the viscosity constant ϵ . Certainly $u_\epsilon, v_\epsilon, p_\epsilon$ tend to u_0, v_0, p_0 with $\epsilon \rightarrow 0$ in the region $z > 0$, but the convergence cannot be uniform at the boundary because the viscous fluid adheres, the ideal glides along the wall. Hence we have the phenomenon of a boundary layer of thickness $\sim \epsilon$ in which the velocity rises from 0 at the surface to the external value

$$\bar{u} = u_0(r, 0) = \frac{2k}{m-1} r, \quad \bar{v} = v_0(r, 0) = 0.$$

Indeed

$$(3) \quad u_\epsilon(r, \epsilon z), \quad \frac{1}{\epsilon} v_\epsilon(r, \epsilon z), \quad p_\epsilon(r, \epsilon z)$$

tend with $\epsilon \rightarrow 0$ to the values

$$U(r, z) = \frac{2}{m-1} r w'(z), \quad V(r, z) = -2w(z),$$

$$\bar{p}(r) = p_0(r, 0) = \text{const.} - 2 \left(\frac{kr}{m-1} \right)^2.$$

[As a matter of fact, the first two quantities (3) are independent of ϵ , the last differs from $\bar{p}(r)$ by the term $2\epsilon^2\{w'(z) + w^2(z)\}$ of order ϵ^2 .] According to L. Prandtl, similar circumstances with regard to convergence for $\epsilon \rightarrow 0$ are to be expected along the surface of any obstacle immersed in a fluid of viscosity ϵ^2 .

We propose to formulate the *two-dimensional* boundary layer problem in terms of *conformal* coordinates ξ_1, ξ_2 which arise from the Cartesian coordinates x_1, x_2 by a conformal transformation. Let u_1, u_2 be the covariant components of velocity with respect to these coordinates ξ_1, ξ_2 and

$$ds^2 = dx_1^2 + dx_2^2 = e(d\xi_1^2 + d\xi_2^2)$$

the square of the line element. The Navier-Stokes equations assume the form

$$(4) \quad \frac{\partial u_1}{\partial \xi_1} + \frac{\partial u_2}{\partial \xi_2} = 0,$$

$$(5) \quad \sum_k \frac{\partial u_i}{\partial \xi_k} u_k - \frac{1}{2} \frac{1}{e} \frac{\partial e}{\partial \xi_i} \sum_k u_k^2 + \frac{\partial p}{\partial \xi_i} = \epsilon^2 \left\{ \Delta u_i + \frac{1}{e} \sum_k \left(\frac{\partial u_k}{\partial \xi_i} - \frac{\partial u_i}{\partial \xi_k} \right) \frac{\partial e}{\partial \xi_k} \right\} \quad [i, k = 1, 2]$$

where

$$\Delta u_i = \frac{\partial^2 u_i}{\partial \xi_1^2} + \frac{\partial^2 u_i}{\partial \xi_2^2}.$$

We suppose that u_1, u_2, p with $\epsilon \rightarrow 0$ converge to the flow \mathcal{F}_0 of an ideal fluid arising from a harmonic potential φ :

$$\overset{0}{u}_i = \partial \varphi / \partial \xi_i, \quad p_0 = \text{const.} - \frac{1}{2} \sum_i \overset{0}{u}_i \overset{0}{u}_i = \text{const.} - \frac{1}{2e} \sum \overset{0}{u}_i^2.$$

Along with φ any multiple $k\varphi$ with a positive constant factor k is equally serviceable, which means that the total strength of the stream may be arbitrarily fixed. Choose ξ_1, ξ_2 so that a multiple $k\zeta$ of $\zeta = \xi_1 + i\xi_2$ is the complex potential of the limiting flow \mathcal{F}_0 and let the stream line $\xi_2 = 0$ be the one which forms the boundary. We have good reasons to believe, and this belief is the basis of the boundary-layer theory, that $u_1, u_2/\epsilon$ and p , when expressed in terms of the arguments $\xi = \xi_1, \eta = \xi_2/\epsilon$, tend to limiting functions $U(\xi, \eta), V(\xi, \eta), P(\xi, \eta)$ which satisfy the equations arising from (4), (5) by the same passage to the limit. The second equation (5) then shows that $\partial P/\partial \eta = 0$, and $P(\xi, \eta)$ is therefore independent of η and has the value

$$\bar{p}(\xi) = p_0(\xi, 0) = \text{const.} - k^2/2\bar{e}(\xi), \quad \bar{e}(\xi) = e(\xi, 0),$$

throughout the boundary layer. Thereafter the two other equations give

$$(B) \quad \begin{cases} \frac{\partial U}{\partial \xi} + \frac{\partial V}{\partial \eta} = 0, \\ U \frac{\partial U}{\partial \xi} + V \frac{\partial U}{\partial \eta} + h(\xi)(k^2 - U^2) = \frac{\partial^2 U}{\partial \eta^2} \end{cases}$$

where

$$h(\xi) = \frac{1}{2} \frac{d \log \bar{e}(\xi)}{d\xi}.$$

One has to add the boundary conditions

$$(\bar{B}) \quad U \rightarrow 0, \quad V \rightarrow 0 \quad \text{for} \quad \eta \rightarrow 0 \quad \text{and} \quad U \rightarrow k \quad \text{for} \quad \eta \rightarrow \infty.$$

A full justification of the basic hypothesis of boundary-layer theory will hardly be possible without changing its differential form as given by these equations into a suitable integral form and without proving the existence of a unique solution of the problem (B, \bar{B}).³

Because of the first equation (B), the flow (U, V) derives from a stream function ψ ,

$$U = \partial\psi/\partial\eta, \quad V = -\partial\psi/\partial\xi$$

satisfying the formidable differential equation

$$(B_\psi) \quad h(\xi) \left\{ k^2 - \left(\frac{\partial\psi}{\partial\eta} \right)^2 \right\} + \frac{\partial^2 \psi}{\partial \xi \partial \eta} \cdot \frac{\partial \psi}{\partial \eta} - \frac{\partial^2 \psi}{\partial \eta^2} \cdot \frac{\partial \psi}{\partial \xi} = \frac{\partial^3 \psi}{\partial \eta^3}$$

and the boundary conditions

$$(\bar{B}_\psi) \quad \psi \rightarrow 0, \quad \frac{\partial \psi}{\partial \eta} \rightarrow 0 \quad \text{for} \quad \eta \rightarrow 0, \quad \frac{\partial \psi}{\partial \eta} \rightarrow k \quad \text{for} \quad \eta \rightarrow \infty.$$

³ Experience shows that in general the assumptions of the theory are fulfilled only along a certain frontal part of the surface of the solid. For the whole theory see S. Goldstein, *Modern Developments in Fluid Dynamics*, vol. I, Oxford, 1938.

Suppose now the obstacle is an angle of $\pi\lambda$ ($0 \leq \lambda < 2$) with the origin as vertex and the positive real axis as median. The exterior of the angle is mapped conformally upon the slit $(\xi_1 + i\xi_2)$ -plane, the slit extending along the positive real axis, by the analytic function

$$x_1 + ix_2 = \text{const. } (\xi_1 + i\xi_2)^{1-\lambda},$$

and thus one readily finds

$$(6) \quad h(\xi) = -\frac{\lambda}{2} \cdot \frac{1}{\xi}.$$

The domain for the differential equation (B_ψ) is the quadrant $\xi > 0$, $\eta > 0$. If, more generally, the solid parts the stream symmetrically with a prow of angle $\pi\lambda$ at the origin, then the formula (6) will hold at least approximately in the neighborhood of the forward stagnation point. The problem (B_ψ , \bar{B}_ψ) with this value of $h(\xi)$ is carried into itself by the transformation

$$(7) \quad \xi \rightarrow \gamma^2 \cdot \xi, \quad \eta \rightarrow \gamma \cdot \eta, \quad \psi \rightarrow \gamma \cdot \psi$$

(γ a positive constant). Hence the solution must be of the form

$$(8) \quad \psi(\xi, \eta) = 2\sqrt{\xi} \cdot w(\eta/2\sqrt{\xi}),$$

and for the function $w(z)$ one obtains exactly the conditions (A_λ). The case $\lambda = 0$ where the obstacle consists of the half line $y = 0$, $x \geq 0$ in the x, y -plane and the fluid flows by with constant positive velocity k was the first boundary-layer problem to be numerically integrated (Blasius 1907). Arbitrary values of λ have been treated by V. M. Falkner, S. W. Skan and D. R. Hartree.⁴ Of the two hydrodynamic interpretations for (A_λ) which we have described, the second is applicable to all values of λ (at least within the range $0 \leq \lambda < 2$), the first to the reciprocal integers $\lambda = 1/(m - 1)$ only. Both coincide for $\lambda = 1$. We notice in particular that the two-dimensional boundary-layer problem of the rectangular prow is mathematically equivalent to the three-dimensional flow against a straight wall ($\lambda = \frac{1}{2}$).

2. Solution of Blasius's Problem

Turning to the solution of our problems, we start with the case $\lambda = 0$, which occupies a singular position inasmuch as the parameter k is absent from its differential equation

$$(9) \quad w''' + 2ww'' = 0.$$

⁴ H. Blasius, *Zeitschr. Math. Phys.* 56, 1908, p. 1. V. M. Falkner and S. W. Skan, *Phil. Mag.* 12, 1931, p. 865; (British) *Aero. Res. Comm. R. & M.* 1314; D. R. Hartree, *Proc. Camb. Phil. Soc.* 33, 1937, pp. 223-239.

For any constant κ the expression $\kappa \cdot w(\kappa z)$ is a solution of this equation if $w(z)$ is. Following an argument first advanced by Töpfer⁵ let $w = f(z)$ be the solution determined by the initial values

$$f(0) = f'(0) = 0, \quad f''(0) = 1.$$

Once we are sure that f extends over the whole interval $0 \leq z < \infty$ and f' tends to a positive limit β with $z \rightarrow \infty$,

$$\beta = \int_0^\infty f''(z) dz > 0,$$

we may adjust the constant κ so as to let the derivative of $w = \kappa \cdot f(\kappa z)$ approach k at infinity:

$$(10) \quad \kappa^2 \beta = k, \quad \kappa = (k/\beta)^{\frac{1}{2}}.$$

Therefore

$$(11) \quad w''(0) = \kappa^3 = \alpha k^{\frac{3}{2}}, \quad \alpha = \beta^{-\frac{3}{2}}.$$

The value $w''(0)$ is the essential factor in the formula for the skin friction along the immersed plate. Hence skin friction is proportional to the $3/2$ power of velocity.

Treat f and f'' in the equation

$$\frac{df''}{dz} + 2ff'' = 0$$

as two separate functions. Because of the initial condition $f''(0) = 1$ one then obtains

$$f''(z) = \exp \left(-2 \int_0^z f(\xi) d\xi \right).$$

Introduce $f'' = g$ as the unknown function and using the initial values $f(0) = f'(0) = 0$, tie up f , or rather its integral, with g by two successive partial integrations:

$$2 \int_0^z f(\xi) d\xi = \int_0^z (z - \xi)^2 \cdot f''(\xi) d\xi.$$

The differential equation plus the initial conditions are thus equivalent to the integral equation

$$(12) \quad g = \Phi\{g\}$$

with the operator

$$\Phi\{g\} = \exp \left(- \int_0^z (z - \xi)^2 g(\xi) d\xi \right).$$

⁵ Zeitschr. Math. Phys. 60, 1912, pp. 397-398.

Notice the following properties of this operator:

$$(i) \quad \Phi\{g\} \geq 0, \quad (ii) \quad \Phi\{g\} \geq \Phi\{g^*\} \quad \text{if } g \leq g^*.$$

We are led to define a sequence of successive "approximations" g_n , starting with $g_0(z) = 0$, by

$$g_{n+1} = \Phi\{g_n\} \quad (n = 0, 1, 2, \dots).$$

The trivial relations $g_1 \geq g_0 = 0$, $g_2 \geq g_0 = 0$, implied in (i), give rise by (ii) to two rows of inequalities, namely

$$(13) \quad g_0 \leq g_1, \quad g_1 \geq g_2, \quad g_2 \leq g_3, \quad g_3 \geq g_4, \dots$$

and

$$(14) \quad g_0 \leq g_2, \quad g_1 \geq g_3, \quad g_2 \leq g_4, \quad g_3 \geq g_5, \dots$$

The latter may be rearranged as follows:

$$g_0 \leq g_2 \leq g_4 \leq \dots \quad \text{and} \quad g_1 \geq g_3 \geq g_5 \geq \dots$$

In view of (13) these relations prove that the descending sequence of the odd g_n lies above the ascending sequence of the even g_n . Does this "alternating pincer movement" close in on a uniquely determined limit function $g(z)$?

To answer this question, introduce the abbreviation

$$(15) \quad G(z) = \int_0^z (z - \zeta)^2 g(\zeta) d\zeta$$

and let

$$0 \leq g(z) \leq g^*(z), \quad \Delta g = g^* - g, \quad \Delta\Phi\{g\} = \Phi\{g^*\} - \Phi\{g\}.$$

Since

$$0 \leq e^{-u} - e^{-v} \leq v - u \quad \text{if } 0 \leq u \leq v$$

we get

$$0 \leq -\Delta\Phi\{g\} \leq \Delta G.$$

The increment ΔG arises from $2 \cdot \Delta g$ by thrice integrating from 0 to z . These remarks suffice to establish the inequality

$$(16) \quad |g_{n+1}(z) - g_n(z)| \leq (2z^3)^n / (3n)!.$$

Indeed, because $g_0 = 0$, $g_1 = 1$, it holds for $n = 0$, and since threefold integration changes

$$z^{3n} / (3n)! \quad \text{into} \quad z^{3n+3} / (3n+3)!$$

the inequality carries over from n to $n+1$, i.e. from $g_{n+1} - g_n$ to $\Phi\{g_{n+1}\} - \Phi\{g_n\}$. Thus convergence (of the type of the exponential series) is assured by the relation (16), and we obtain a solution

$$g(z) = \lim_{n \rightarrow \infty} g_n(z)$$

of (12) which is larger than the even and smaller than the odd g_n .

Uniqueness is established by the remark that *any* solution $g(z)$ of (12) satisfies the inequalities

$$g \geq g_0, \quad g \leq g_1, \quad g \geq g_2, \quad g \leq g_3, \dots$$

derived from the trivial one $g \geq g_0 = 0$ by iterated application of the operator Φ . Thus g is necessarily caught between the tongs of the even and the odd g_n .

Our next concern is the asymptotic behavior of $g(z)$ for $z \rightarrow \infty$. Choose any $z_0 > 0$ and set

$$\int_0^{z_0} g_2(\xi) d\xi = c (> 0).$$

As G_2 arises by two-fold integration from $2 \int_0^z g_2(\xi) d\xi$ we get $G_2(z) \geq c(z - z_0)^2$ and thus

$$g(z) \leq g_3(z) \leq e^{-c(z-z_0)^2} \quad \text{for } z \geq z_0.$$

Consequently

$$\int_0^\infty g(z) dz = \beta > 0, \quad \int_0^\infty zg(z) dz = \beta' > 0$$

converge, the asymptotic behavior of

$$f(z) = \int_0^z (z - \xi)g(\xi) d\xi$$

is indicated by

$$(17) \quad f(z) \sim \int_0^\infty (z - \xi)g(\xi) d\xi = \beta z - \beta'$$

and that of $w(z)$ by

$$(18) \quad w(z) \sim kz - \frac{\beta'}{\beta^{\frac{1}{2}}} k^{\frac{1}{2}}.$$

As $g(z) \geq g_2(z) = e^{-\frac{1}{2}z^2}$ implies

$$\beta \geq B = \int_0^\infty e^{-\frac{1}{2}z^2} dz = 3^{\frac{1}{2}} \cdot \Gamma\left(\frac{4}{3}\right)$$

we find the numerical coefficient α in (11) to be < 0.684 . According to the most reliable computations⁶ $\alpha = 0.664$. Hence the very first approximate value for α which can be derived from our method misses the mark by not more than 3 per cent.

Given any positive constant $c < B$ we have seen that, for sufficiently large z ,

$$(G_n(z) \geq) G_2(z) \geq cz^2.$$

⁶ See Töpfer, l.c.⁵, and S. Goldstein, Proc. Camb. Phil. Soc. 26, 1930, pp. 19-20.

By making use of this fact, one can sharpen the upper bound in (16) to

$$e^{-cs^3} \cdot (2z^3)^n / (3n)!$$

The maximum value of this function is assumed for $z = (3n/2c)^{1/3}$, and we thus ascertain a constant upper bound μ_n for $|g_{n+1}(z) - g_n(z)|$ in the entire interval $z \geq 0$ which is essentially of the order

$$\frac{1}{\sqrt{(6\pi n)}} (\mu n)^{-3n/2}, \quad \mu \sim 1.8.$$

Such sharper estimates are valuable guides for numerical computation.

Knowing a priori that the positive functions f'' , f' , f have the upper bounds 1, z , $\frac{1}{2}z^2$ respectively, one could have established the existence (and uniqueness) of the solution f over the entire interval $0 \leq z < \infty$ within the frame of the classic theory of differential equations. But as those bounds (and other related estimates) are most easily derived from the integral equation (12), I have preferred to carry the construction through on its basis. For the general case (A_λ), $\lambda \neq 0$, I see no other alternative.

J. von Neumann pointed out to me that the differential equation (9) of order 3 must be reducible to one of first order (followed by two quadratures) because it permits the group of transformations

$$z \rightarrow z + z_0, \quad w(z) \rightarrow \kappa \cdot w(\kappa z)$$

involving two arbitrary constants z_0 and κ , and that thus the problem comes within reach of Poincaré's discussion of first-order differential equations. Setting

$$w = e^{-s}, \quad \frac{dw}{dz} = e^{-2s} \cdot \vartheta(s), \quad -\frac{d\vartheta}{ds} + 2\vartheta = t$$

von Neumann obtains the equation

$$\frac{dt}{d\vartheta} = \frac{t(t + \vartheta + 2)}{\vartheta(2\vartheta - t)}$$

with the initial condition $t \rightarrow \infty$ for $\vartheta \rightarrow \infty$. After determining $t(\vartheta)$ from this equation, one finds by quadratures s and then z as functions of ϑ from

$$ds = \frac{d\vartheta}{2\vartheta - t}, \quad dz = -\frac{e^s ds}{\vartheta} = -e^s \cdot \frac{d\vartheta}{\vartheta(2\vartheta - t)}.$$

3. Generalization. Power series. Goldstein's wake problem.

In a trivial manner our method carries over to the equation

$$(19) \quad f^{(r+1)} + 2ff^{(r)} = 0 \quad (z \geq 0)$$

with the initial conditions

$$(20) \quad f = f' = \dots = f^{(r-1)} = 0, \quad f^{(r)} = 1 \quad \text{for } z = 0.$$

Here ν may be any positive integer. Setting $f^{(\nu)} = g$ we get the integral equation

$$g(z) = \exp \left(-\frac{2}{\nu!} \int_0^z (z' - \zeta)^\nu g(\zeta) d\zeta \right),$$

and after defining the alternating sequence $g_n(z)$ accordingly, we find instead of (16)

$$|g_{n+1}(z) - g_n(z)| \leq 2^n z^{n^*} / n^*! \quad [n^* = (\nu + 1)n].$$

We may even generalize the initial conditions (20) to

$$(21) \quad f(0) = c_0, \dots, f^{(\nu-1)}(0) = c_{\nu-1}, \quad f^{(\nu)}(0) = 1$$

with arbitrary constants c_μ . Then our integral equation reads

$$g(z) = \exp \left(-2Q(z) - \frac{2}{\nu!} \int_0^z (z - \zeta)^\nu g(\zeta) d\zeta \right),$$

where $Q(z)$ is the polynomial

$$(22) \quad Q(z) = \frac{c_0}{1!} z + \frac{c_1}{2!} z^2 + \dots + \frac{c_{\nu-1}}{\nu!} z^\nu,$$

and convergence follows from the inequality

$$|g_{n+1}(z) - g_n(z)| \leq A^{n+1} \cdot 2^n z^{n^*} / n^*! [\leq A(2Aa^{\nu+1})^n / n^*!]$$

holding in any interval $0 \leq z \leq a$ in which $e^{-2Q(z)} \leq A$. The solution $g(z)$ satisfies the inequality

$$(23) \quad 0 \leq g(z) \leq g_1(z) = e^{-2Q(z)}.$$

Let us for a moment return to the simple initial conditions (20). From the lowest case $\nu = 1$ where the solution is an elementary function, namely $f(z) = \tanh(z)$, we learn that we must not expect the Taylor expansion of the solution $f(z)$ around the origin to converge beyond a certain finite limit, which for $\nu = 1$ is reached at the point $z = \pi/2$. I find no indication in the literature that this had been realized in Blasius's case $\nu = 2$. For any ν the coefficients c_n of the power series

$$f(z) = \sum_{n=0}^{\infty} (-1)^n c_n z^{n^*+\nu} \quad [n^* = (\nu + 1)n]$$

are determined by the recursive equations $c_0 = 1/\nu!$,

$$n^*(n^* + 1) \cdots (n^* + \nu) c_n = 2 \sum (i^* + 1) \cdots (i^* + \nu) c_i c_k \quad (i + k = n - 1).$$

Following the same straightforward procedure as in my first note in the Proceedings, we obtain

$$\frac{1}{\nu!} \left\{ \frac{2 \cdot \nu!}{(2\nu + 1)!} \right\}^n \leq c_n \leq \frac{1}{\nu!} \left\{ \frac{2}{(\nu + 1) \cdot (\nu + 1)!} \right\}^n$$

and thus for the radius R of convergence the bounds

$$\frac{1}{2}(\nu + 1) \cdot (\nu + 1)! \leq R^{\nu+1} \leq \frac{1}{2}(\nu + 1) \cdots (2\nu + 1).$$

The essential difference between the flow of a viscous fluid *before* and *behind* an obstacle is clearly exhibited in our problem (A_0) by the fact that no solution w exists if w' is required to assume a *negative* value k for $z \rightarrow \infty$.

However, S. Goldstein⁷ has treated the wake behind a flat plate under the plausible hypothesis that the flow up to the abscissa $x = l$ is but little modified if the plate, $y = 0$, $x \geq 0$, ends at this point. We shift the origin of the coordinates to the end of the plate and at the same time enlarge the standard length at the ratio $l^{\frac{1}{2}}:1$, i.e. in our stream function

$$\psi(x, y) = 2\sqrt{x} \cdot w(y/2\sqrt{x})$$

we make the substitution

$$x = l + l^{\frac{1}{2}} \cdot \xi, \quad y = l^{\frac{1}{2}} \cdot \eta.$$

We then obtain for $\xi = 0$:

$$\psi = \varphi(\eta) + \cdots, \quad \varphi(\eta) = \frac{1}{4}\alpha k^{\frac{1}{2}} \cdot \eta^2.$$

The remainder indicated by the dots tends to zero with $l \rightarrow \infty$ and shall be neglected as is permissible for plates of great length. The stream function $\psi(\xi, \eta)$ of the "wake layer" behind the plate, $\xi > 0$, satisfies the same differential equation as before

$$\frac{\partial^3 \psi}{\partial \eta^3} + \frac{\partial \psi}{\partial \xi} \cdot \frac{\partial^2 \psi}{\partial \eta^2} - \frac{\partial \psi}{\partial \eta} \cdot \frac{\partial^2 \psi}{\partial \xi \partial \eta} = 0$$

while symmetry requires $\psi(\xi, -\eta) = -\psi(\xi, \eta)$. Hence under limitation to the half plane $\eta \geq 0$ the conditions at the fictitious boundary $\eta = 0$ become $\psi = \partial^2 \psi / \partial \eta^2 = 0$. We wish to construct that solution which for fixed η and $\xi \rightarrow 0$ (or for fixed ξ and $\eta \rightarrow \infty$) ties up with our function $\varphi(\eta)$. The problem, including this boundary condition, permits the substitution

$$\xi \rightarrow \gamma^3 \cdot \xi, \quad \eta \rightarrow \gamma \cdot \eta, \quad \psi \rightarrow \gamma^2 \cdot \psi$$

with an arbitrary constant γ and must thus be of the form

$$\psi(\xi, \eta) = 3\xi^{\frac{1}{2}} \cdot w(\eta/\xi^{\frac{1}{2}}).$$

For $w(z)$ one readily obtains the differential equation

$$(24) \quad w''' + 2ww'' - w'^2 = 0.$$

The boundary conditions are: $w = w'' = 0$ at $z = 0$, and

$$(25) \quad w''(z) \rightarrow \frac{1}{4}\alpha k^{\frac{1}{2}} \quad \text{for } z \rightarrow \infty.$$

⁷ Proc. Camb. Phil. Soc. 26, 1930, pp. 18-30.

If $w(z)$ is a solution of (24), so is the function $\kappa \cdot w(\kappa z)$ involving an arbitrary constant κ . Let $w = f(z)$ be the solution of (24) with the initial values $f = 0$, $f' = 1$, $f'' = 0$ for $z = 0$. Then $f'''(0) = 1$ while differentiation changes (24) into

$$(26) \quad w'''' + 2ww''' = 0.$$

Thus we find ourselves confronted with the case $\nu = 3$; $c_0 = 0$, $c_1 = 1$, $c_2 = 0$ of the general problem (19) + (21) discussed above, and since (23) now reads

$$0 \leq f'''(z) = g(z) \leq e^{-z^2},$$

$f'''(z)$ tends with $z \rightarrow \infty$ to a positive limit

$$\beta^* = \int_0^\infty g(z) dz < \sqrt{\frac{\pi}{2}}.$$

Consequently we may adjust the constant κ^* in $w(z) = \kappa^* \cdot f(\kappa^* z)$ so as to give $w''(\infty)$ the desired value (25).^{7a}

4. Solution of Homann's problem

Of a more difficult type is the problem (A_λ) for $\lambda \neq 0$, as it involves the parameter k in the boundary conditions as well as in the differential equation itself. Here we are dealing with a real boundary-value problem, which is not reducible, as (A_0) is, to an initial-value problem. Only convergence of the type of a geometric series if any can be expected for the process of successive approximations. By differentiating the differential equation we eliminate the parameter k :

$$(27) \quad w'''' + 2ww''' + 2(1 - 2\lambda)w'w'' = 0.$$

This equation is again invariant under the transformation $w(z) \rightarrow \kappa \cdot w(\kappa z)$. For $\lambda = \frac{1}{2}$, the case with which we shall be concerned in the next two sections, we fall back upon the familiar type (26), although the boundary conditions make our problem considerably more intricate than before. Let f denote that solution of (26) for which

$$f(0) = f'(0) = 0, \quad f''(0) = 1, \quad f'''(0) = -\beta^2.$$

It will satisfy the third-order equation

$$f''' + 2ff'' + (\beta^2 - f'^2) = 0,$$

and we expect that, for a certain positive β , the derivative f'' (and f''') will strongly approach 0 with $z \rightarrow \infty$. Thus the equation itself forces f' (which is positive throughout the interval) to approach β , and $w = \kappa \cdot f(\kappa z)$ will solve our problem if κ is determined by (10).

^{7a} A remark by K. Friedrichs to the effect that the assumptions $\nu = 3$, $2Q(z) = -z^2$ lead to a wake with back flow caused me to drop the restriction $Q(z) \geq 0$ for $z \geq 0$ which the original MS contained. April 4, 1948.

As before we obtain first

$$\begin{aligned} f'''(z) &= -\beta^2 \cdot \exp \left(-2 \int_0^z f(\xi) d\xi \right) \\ &= -\beta^2 \cdot \exp \left(- \int_0^z (z - \xi)^2 f''(\xi) d\xi \right) = -\beta^2 \cdot e^{-g(z)} \end{aligned}$$

and then for $f'' = g$ the equation

$$g(z) = 1 - \beta^2 \cdot \int_0^z e^{-g(\xi)} d\xi.$$

The constant β^2 is determined by the condition $g(\infty) = 0$, thus

$$\beta^2 = 1 / \int_0^\infty e^{-g(\xi)} d\xi.$$

Adhering to the notation (15) we are led to introduce an operator Ψ which produces from any given $g(z)$ the function

$$(28) \quad \Psi\{g\} = \int_z^\infty e^{-g(\xi)} d\xi / \int_0^\infty e^{-g(\xi)} d\xi$$

(provided the integral \int_0^∞ converges). Evidently

$$0 < \Psi\{g\} \leq 1,$$

and as we shall presently prove,

$$(29) \quad \Psi\{g\} \geq \Psi\{g^*\} \quad \text{if } g \leq g^*.$$

The operator Ψ is applicable to the function $g(z) = 1$ but not to $g(z) = 0$.⁸ In order to solve the functional equation

$$(30) \quad g = \Psi\{g\}$$

we therefore construct a sequence of functions $g_n(z)$ by the recursive equation

$$g_{n+1} = \Psi\{g_n\} \quad (n = 1, 2, \dots)$$

starting with $g_1(z) = 1$, in the hope that the sequence will converge to a solution g of (30). Alternating pincer movement of the g_n is a consequence of (29) and the trivial inequalities $g_2 \leq g_1$, $g_3 \leq g_1$.

To prove (29), set as before $g^* - g = \Delta g$. The third derivative of $\Delta G = G^* - G$ is $2 \cdot \Delta g$, and ΔG and its first two derivatives vanish for $z = 0$. Hence

⁸ Application of the operator Ψ becomes unrestricted if one replaces the definition (28) by

$$\Psi\{g\} = \lim_{a \rightarrow \infty} \left(\int_z^a / \int_0^a \right).$$

Then one may start with $g_0(z) = 0$ and find $g_1(z) = 1$. Cf. §6.

$\Delta g \geq 0$ implies $(\Delta G)''$ to be an increasing function of z , and as it vanishes for $z = 0$ it must be positive throughout. Repeating this argument two more times, we find that $\Delta G(z)$ is an increasing positive function for $z > 0$. Set

$$\int_0^z e^{-G(\zeta)} d\zeta = H_1, \quad \int_z^\infty e^{-G(\zeta)} d\zeta = H_2,$$

so that

$$\Psi\{g\} = H_2/(H_1 + H_2).$$

We then have

$$H_1^* = \int_0^z e^{-G^*(\zeta)} d\zeta = \int_0^z e^{-G(\zeta)} \cdot e^{-\Delta G(\zeta)} d\zeta \geq e^{-\Delta G(z)} \cdot H_1,$$

$$H_2^* = \int_z^\infty e^{-G^*(\zeta)} d\zeta = \int_z^\infty e^{-G(\zeta)} \cdot e^{-\Delta G(\zeta)} d\zeta \leq e^{-\Delta G(z)} \cdot H_2,$$

or the ratios $\vartheta_1 = H_1^*/H_1$, $\vartheta_2 = H_2^*/H_2$ satisfy the inequalities ($\vartheta_1 \leq 1$, $\vartheta_2 \leq 1$), $\vartheta_2 \leq \vartheta_1$. Consequently

$$H_2^*/(H_1^* + H_2^*) \leq H_2/(H_1 + H_2) \quad \text{or} \quad \Psi\{g^*\} \leq \Psi\{g\}.$$

Again choose a $z_0 > 0$ and set

$$\int_0^{z_0} g_2(\zeta) d\zeta = c > 0$$

so that $G_2(z) \geq c(z - z_0)^2$ and

$$g_3(z) \leq \text{const.} \int_z^\infty e^{-c(\zeta - z_0)^2} d\zeta \quad \text{for } z \geq z_0.$$

All following g 's are smaller than g_3 and hence the same inequality prevails for $g_n(z)$ ($n \geq 3$) and, provided the limit $\lim_{n \rightarrow \infty} g_n(z) = g(z)$ exists, also for $g(z)$. Thus knowing that $g(z) = f''(z)$ converges strongly enough to 0 with $z \rightarrow \infty$ we again get the asymptotic formula (17) and (11), (18) for the solution

$$w(z) = \kappa \cdot f(\kappa z), \quad \kappa = (k/\beta)^{\frac{1}{2}}$$

of our problem (A_1).

In proving *convergence* of the alternating sequence $g_n(z)$ we use the above notations $g(z) \leq g^*(z)$, Δg , G etc., and write $H = H_1 + H_2$. Then

$$\begin{aligned} -\Delta\Psi &= \Psi\{g\} - \Psi\{g^*\} = \frac{H_2}{H} - \frac{H_2^*}{H^*} \\ &= \frac{H_2 - H_2^*}{H} + H_2^* \left(\frac{1}{H} - \frac{1}{H^*} \right) \leq \frac{H_2 - H_2^*}{H} \leq \frac{H - H^*}{H}. \end{aligned}$$

Thus

$$-\Delta\Psi \leq \int_0^\infty e^{-G(\zeta)} (1 - e^{-\Delta G(\zeta)}) d\zeta \bigg/ \int_0^\infty e^{-G(\zeta)} d\zeta.$$

Suppose we have a constant μ such that

$$0 \leq \Delta g(z) \leq \mu.$$

Then

$$1 - e^{-\Delta g(\xi)} \leq \Delta G(\xi) \leq \mu \cdot \frac{1}{3} \xi^3$$

and hence

$$0 \leq -\Delta \Psi \leq q \cdot \mu$$

where

$$q = \int_0^\infty e^{-G(z)} \cdot \frac{1}{3} z^3 dz \bigg/ \int_0^\infty e^{-G(z)} dz.$$

Another expression for the quotient q is

$$\int_0^\infty \Psi\{g\} \cdot z^2 dz$$

as one verifies by substituting (28) for $\Psi\{g\}$.

In this argument we can choose $g = g_n$ and $g^* = g_{n+1}$ or g_{n-1} for any even $n \geq 2$ and then we obtain majorizing constants μ_n for all odd and even $n \geq 1$,

$$(31) \quad |g_{n+1}(z) - g_n(z)| \leq \mu_n,$$

which are defined by the recursive equations

$$(32) \quad \mu_1 = 1; \quad \mu_n = q_n \cdot \mu_{n-1}, \quad \mu_{n+1} = q_n \cdot \mu_n \quad (n \text{ even})$$

with

$$\begin{aligned} q_n &= \int_0^\infty e^{-G_n(z)} \cdot \frac{1}{3} z^3 dz \bigg/ \int_0^\infty e^{-G_n(z)} dz \\ &= \int_0^\infty g_{n+1}(z) \cdot z^2 dz. \end{aligned}$$

The second expression of q_n shows that the constants q_n perform a pincer movement of the same type as the functions $g_n(z)$, and hence all q_n lie between q_1 and q_2 . Evaluating by partial integration the integral in the numerator of

$$q_1 = \int_0^\infty e^{-\frac{1}{3}z^3} \cdot \frac{1}{3} z^3 dz \bigg/ \int_0^\infty e^{-\frac{1}{3}z^3} dz,$$

namely

$$-\frac{1}{3} \int_0^\infty z \cdot de^{-\frac{1}{3}z^3},$$

we find $q_1 = \frac{1}{3}$. By some rough estimates it is proved in §5 that $q_2 < 0.76$; but the value of q_2 is probably not much larger than $q_1 = 0.33$. Once we are sure that $q_2 < 1$ we see from (32) and $q_n \leq q_2$ that the sequence $g_n(z)$ converges at least as strongly as a geometric series of quotient q_2 .

Uniqueness is assured since every solution g of the equation $g = \Psi\{g\}$ is necessarily sandwiched in between the odd and even g_n .

5. Proof that $q_2 < 1$

We use the constants

$$A = \int_0^{\infty} z \cdot e^{-iz^3} dz = 3^{-1} \cdot \Gamma(\frac{2}{3}),$$

$$B = \int_0^{\infty} e^{-iz^3} dz = 3^{-1} \cdot \Gamma(\frac{1}{3}) = 1.288.$$

Their product is

$$AB = \frac{1}{3} \Gamma(\frac{1}{3}) \Gamma(\frac{2}{3}) = \frac{\pi}{3 \sin(\pi/3)} = \frac{2\pi}{3\sqrt{3}}.$$

q_2 is defined as the quotient

$$\int_0^{\infty} e^{-G_2(z)} \cdot \frac{1}{3} z^3 dz \bigg/ \int_0^{\infty} e^{-G_1(z)} dz.$$

Since

$$G_2(z) < G_1(z) = \frac{1}{3} z^3$$

the denominator is greater than B . Let us split the integral of the numerator into the parts

$$\int_0^2 + \int_2^{\infty}$$

and employ in the first part the initial terms of the power series of $G_2(z)$, in the second part an asymptotic appraisal. We readily find

$$(33) \quad G_2(z) = \frac{1}{3} z^3 - \frac{1}{B} \int_0^z \frac{1}{3} (z - \zeta)^3 \cdot e^{-i\zeta^3} d\zeta$$

and, because $e^{-i\zeta^3} \leq 1$,

$$G_2(z) > \frac{1}{3} z^3 - \frac{1}{12B} z^4$$

and thus, as long as $z \leq 2$,

$$G_2(z) > c \cdot \frac{1}{12} z^4 \quad \text{with} \quad c = 2 - 1/B.$$

Consequently

$$\begin{aligned} \int_0^2 e^{-G_2(z)} \cdot \frac{1}{3} z^3 dz &< \int_0^2 e^{-c z^4/12} \cdot d \frac{z^4}{12} \\ &= \frac{1}{c} (1 - e^{-4c/3}) = 0.6574. \end{aligned}$$

To find an asymptotic estimate write (33) in the form

$$B \cdot G_2(z) = \int_0^\infty \frac{1}{3}[z^3 - (z - \zeta)^3] \cdot e^{-\frac{1}{3}\zeta^3} d\zeta - \frac{1}{3} \int_z^\infty (\zeta - z)^3 \cdot e^{-\frac{1}{3}\zeta^3} d\zeta.$$

The first term

$$= Az^2 - z + \frac{1}{3}B,$$

the integral of the second term is changed by the substitution $\zeta \rightarrow \zeta + z$ into

$$\begin{aligned} \int_0^\infty \zeta^3 \cdot e^{-\frac{1}{3}(\zeta+z)^3} d\zeta &\leq e^{-\frac{1}{3}z^3} \cdot \int_0^\infty \zeta^3 \cdot e^{-\frac{1}{3}\zeta^3} d\zeta \\ &= 6z^{-3} e^{-\frac{1}{3}z^3}. \end{aligned}$$

Hence

$$(34) \quad G_2(z) \geq \frac{A}{B} z^2 - \frac{1}{B} z + \left(\frac{1}{3} - \frac{1}{128B} e^{-8/3} \right) \quad \text{for } z \geq 2.$$

Set

$$2\sqrt{AB} = 1/b, \quad \text{i.e.,} \quad (2b)^4 = 27/(2\pi)^2;$$

$$\frac{z}{2Bb} - b = x, \quad \frac{1}{Bb} - b = x_0, \quad \frac{1}{3} - b^2 - \frac{1}{128B} e^{-8/3} = B'$$

so that the right side of (34) equals $x^2 + B'$. Then

$$\begin{aligned} \int_2^\infty &\leq \int_2^\infty e^{-(x^2+B')} \cdot \frac{1}{3} z^3 dz \\ &= \frac{1}{3} (2Bb)^4 e^{-B'} \cdot \int_{x_0}^\infty e^{-x^2} (x+b)^3 dx. \end{aligned}$$

Developing

$$(x+b)^3 = x^3 + 3x^2b + 3xb^2 + b^3$$

one readily finds

$$\int_{x_0}^\infty e^{-x^2} (x+b)^3 dx = \frac{1}{2} e^{-x_0^2} (x_0^2 + 1 + 3bx_0 + 3b^2) + b(\frac{3}{2} + b^2) \int_{x_0}^\infty e^{-x^2} dx.$$

But

$$\int_{x_0}^\infty e^{-x^2} dx = \frac{1}{2} \int_{x_0^2}^\infty e^{-t} \cdot \frac{dt}{\sqrt{t}} \leq \frac{1}{2x_0} \int_{x_0^2}^\infty e^{-t} dt = \frac{1}{2x_0} e^{-x_0^2}.$$

Hence

$$\begin{aligned} \int_2^\infty &\leq \frac{1}{3} (2Bb)^4 e^{-(B'+x_0^2)} \cdot \left\{ x_0^2 + 1 + 3bx_0 + 3b^2 + \frac{b}{x_0} \left(\frac{3}{2} + b^2 \right) \right\} \\ &= 0.3171. \end{aligned}$$

The numerator \int_0^∞ of q_2 turns out to be

$$< 0.6574 + 0.3171 = 0.9745$$

and q_2 itself

$$< 0.9745/B < 0.76.$$

6. Set-up for arbitrary λ

Enriched by the experience gathered in the cases $\lambda = 0$ and $\frac{1}{2}$ we now make bold to attack (A_λ) for arbitrary $\lambda \geq 0$. We seek the solution in the form $w(z) = \kappa \cdot f(\kappa z)$ where

$$(35) \quad \begin{cases} f'''' + 2ff'''' + 2(1 - 2\lambda)f'f'' = 0 & (\text{for } z \geq 0); \\ f(0) = f'(0) = 0, \quad f''(0) = 1; \quad f''(\infty) = 0, \end{cases}$$

and introduce $f'' = g = \varphi$ as the unknown function. Hence we start with this set-up:

$$(36) \quad f(z) = \int_0^z (z - \zeta)g(\zeta) d\zeta.$$

$$(37) \quad \begin{cases} \varphi'' + 2f\varphi' + 2(1 - 2\lambda)f'\varphi = 0, \\ \varphi(0) = 1, \quad \varphi(\infty) = 0. \end{cases}$$

$$(37^*) \quad \begin{cases} \varphi'' + 2f\varphi' + 2(1 - 2\lambda)f'\varphi = 0, \\ \varphi(0) = 1, \quad \varphi(\infty) = 0. \end{cases}$$

$$(38) \quad \varphi = g.$$

More explicitly: for an arbitrarily given function g we form (36) and then solve the linear boundary value problem (37) + (37*), thus defining the functional operator Φ_λ carrying g into φ ; at the last step (38) we ask for a fixed element g of that operator,

$$(39) \quad g = \Phi_\lambda\{g\}.$$

In proving the unique existence of φ we shall fix the precise meaning of the boundary condition $\varphi(\infty) = 0$. (The whole discussion would turn out a bit simpler if we dealt with a finite interval $0 \leq z \leq a$ instead.)

AUXILIARY THEOREM. *If $g(z)$ is any continuous non-negative function and*

$$f'(z) = \int_0^z g(\zeta) d\zeta, \quad f(z) = \int_0^z f'(\zeta) d\zeta = \int_0^z (z - \zeta)g(\zeta) d\zeta$$

then (37) has a unique solution with the properties

$$(40) \quad \varphi(0) = 1; \quad \varphi(z) \geq 0, \quad \varphi' + 2f\varphi \leq 0 \quad (\text{for } z \geq 0).$$

PROOF. Set

$$p(z) = \exp \left(2 \int_0^z f(\zeta) d\zeta \right)$$

so that $p' = 2fp$ and $p(z) \geq 1$, and introduce the auxiliary function

$$\varphi_1 = (p\varphi)' = p(\varphi' + 2f\varphi).$$

Then

$$(\varphi_1/p)' = \varphi'' + 2f\varphi' + 2f'\varphi = 4\lambda f'\varphi,$$

and the single equation (37) for φ is replaced by the system

$$(41) \quad \begin{cases} \varphi' + 2f\varphi = \frac{1}{p} \cdot \varphi_1, \\ \varphi_1' - 2f\varphi_1 = 4\lambda pf' \cdot \varphi. \end{cases}$$

It defines an infinitesimal linear transformation in two variables φ, φ_1 with the matrix (of vanishing trace)

$$\Theta(z) = \begin{vmatrix} -2f & 1/p \\ 4\lambda pf' & 2f \end{vmatrix}.$$

Hence the two solutions $(\varphi, \varphi_1) = (\eta, \eta_1)$ and (ϑ, ϑ_1) satisfying the initial conditions

$$\eta = 1, \quad \eta_1 = 0; \quad \vartheta = 0, \quad \vartheta_1 = 1 \quad (\text{for } z = 0)$$

are given by the formula

$$(42) \quad \begin{vmatrix} \eta & \vartheta \\ \eta_1 & \vartheta_1 \end{vmatrix} = \sum_{n=0}^{\infty} \int \cdots \int_{\substack{0 \leq z_1 \leq \cdots \\ \leq z_n \leq z}} \Theta(z_1) \cdots \Theta(z_n) dz_1 \cdots dz_n$$

(where the term $n = 0$ of the series at the right is understood to be the unit matrix). Multiplication of (41) by φ_1, φ respectively, followed by addition and integration, establishes the fundamental relation

$$(43) \quad [\varphi\varphi_1]_a^b = \int_a^b \left(\frac{1}{p} \varphi_1^2 + 4\lambda pf' \varphi^2 \right) dz.$$

The fact that $f' \geq 0$ guarantees the positive definite character of the "Dirichlet integral" at the right side.

Apply (43) to ϑ :

$$\vartheta\vartheta_1 = \int_0^z \left(\frac{1}{p} \vartheta_1^2 + 4\lambda pf' \vartheta^2 \right) dz > 0 \quad \text{for } z > 0.$$

This shows (1) that ϑ_1 never vanishes, therefore never changes sign and, because of $\vartheta_1(0) = 1$, stays positive throughout; and (2) that ϑ has the same sign as ϑ_1 . In the same manner we find that $\eta\eta_1, \eta, \eta_1$ (in this order) are all positive,

$$\vartheta > 0, \quad \vartheta_1 > 0; \quad \eta > 0, \quad \eta_1 > 0 \quad \text{for } z > 0.$$

Next consider a finite interval $0 \leq z \leq a$ ($a > 0$) and determine that solution φ for which $\varphi(a) = 0$, $\varphi_1(a) = -1$. Again we see from

$$\varphi\varphi_1 = -\int_z^a \left(\frac{1}{p} \varphi_1^2 + 4\lambda p f' \varphi^2 \right) dz < 0$$

that φ_1 is negative and φ positive throughout the interval $0 \leq z < a$. In particular, $\varphi(0) > 0$, so that we can divide by $\varphi(0)$ thus constructing a solution $\varphi^{(a)}$ with the boundary values

$$\varphi^{(a)}(0) = 1, \quad \varphi^{(a)}(a) = 0.$$

It satisfies the inequalities

$$\varphi^{(a)} > 0, \quad \varphi_1^{(a)} < 0 \quad \text{for } 0 \leq z < a.$$

Clearly $\varphi^{(a)}(z)$ is of the form $\eta(z) - l_a \cdot \vartheta(z)$ with a constant l_a for which we find the *positive* value $\eta(a)/\vartheta(a)$. Let $a < b$ and write

$$\varphi^{(b)}(z) = \varphi^{(a)}(z) + (l_a - l_b)\vartheta(z).$$

Then

$$l_a - l_b = \varphi^{(b)}(a)/\vartheta(a) > 0,$$

or the positive coefficient l_a decreases with increasing a and thus tends to a limit $l \geq 0$ for $a \rightarrow \infty$. The solution

$$\omega(z) = \eta(z) - l \cdot \vartheta(z)$$

is the one we wish to construct.⁹ It has the properties

$$(44) \quad \omega(0) = 1; \quad \omega(z) > 0, \quad \omega_1(z) \leq 0,$$

and is characterized by the fact that the condition

$$\varphi(z) = \omega(z) - m \cdot \vartheta(z) \geq 0$$

cannot be satisfied throughout the interval $0 \leq z < \infty$ for any positive constant m .

It remains to show that no solution φ except this ω satisfies (40). Indeed, according to what has just been stated, any such solution would have to be of the form

$$\varphi(z) = \omega(z) + m \cdot \vartheta(z), \quad m > 0.$$

The required inequality $\varphi_1 \leq 0$ or $(p\varphi)' \leq 0$ implies $p\varphi \leq 1$ for $z \geq 0$. This remarkable relation prevails in particular for $\varphi = \omega$. On the other hand,

$$(\vartheta_1/p)' = 4\lambda f' \vartheta \geq 0,$$

⁹ A similar construction in H. Weyl, *Nachr. Ges. Wissensch. Göttingen*, 1909, p. 39.

therefore

$$\begin{aligned} \vartheta_1/p &\geq 1, & \vartheta_1 &\geq p; \\ (p\vartheta)' &= \vartheta_1 \geq p, & p\vartheta &\geq \int_0^x p(\zeta) d\zeta \geq z. \end{aligned}$$

The consequent relation $p\varphi \geq m \cdot z$ is incompatible with $p\varphi \leq 1$ for positive m .

We have now completely and unambiguously defined the functional operator Φ_λ carrying a given $g(z) \geq 0$ into the function $\omega(z)$, $\omega = \Phi_\lambda\{g\}$. Since

$$(45) \quad p\omega \leq 1, \quad \text{a fortiori} \quad \omega \leq 1,$$

our operator Φ_λ obeys the law

$$0 < \Phi_\lambda\{g\} \leq 1.$$

Hence we can and will restrict ourselves to the set \mathcal{G} of all continuous functions $g(z)$ for which $0 \leq g \leq 1$. Were Φ_λ monotone in the sense that $\Phi_\lambda\{g\}$ decreases while g increases, then there would be some hope for successful construction of a fixed point of the operator Φ_λ in the functional space \mathcal{G} by some such alternating process of successive approximation as carried us through in the special instances $\lambda = 0$ and $\frac{1}{2}$. Unfortunately this does not seem to be so, and this calamity forces me to proceed by the general theory concerning fixed points of functional operators which we owe to Birkhoff-Kellogg and Schauder-Leray.¹⁰ The main point will be to establish "equi-continuity" for the images $\omega = \Phi_\lambda\{g\}$ of all elements $g \in \mathcal{G}$ and continuity for the operator Φ_λ . The lemmas in the following section are so conceived as to meet this demand.

7. Solving the problem (A_λ)

In the lemmas 1-4 the function g is supposed to be any element of \mathcal{G} and c, c_0, c_1, c_2, c_3 are numbers not depending on g . The condition $g \leq 1$ implies $f' \leq z$.

LEMMA 1.

$$0 < -\omega'(0) \leq c_1.$$

PROOF. Denote $-\omega'(0)$ by l and argue as follows:

$$\begin{aligned} (\omega_1/p)' &= 4\lambda f'\omega \leq 4\lambda z, \\ \omega_1/p &\leq -l + 2\lambda z^2, \end{aligned}$$

and then, because $p \geq 1$ and ω_1 negative,

$$\begin{aligned} (p\omega)' &= \omega_1 \leq \omega_1/p \leq -l + 2\lambda z^2, \\ p\omega &\leq 1 - lz + \frac{2}{3}\lambda z^3. \end{aligned}$$

¹⁰ G. D. Birkhoff and O. D. Kellogg, Trans. Am. Math. Soc. 23, 1922, pp. 96-115. J. Schauder, Studia Math. 2, 1930, pp. 171-180. J. Leray and J. Schauder, Ann. Sc. Ec. Norm. Sup. 51, 1934, pp. 45-78.

Since $\omega(z) > 0$ we get

$$1 - lz + \frac{2}{3}\lambda z^3 > 0 \quad \text{or} \quad l < \frac{1}{z} + \frac{2}{3}\lambda z^2,$$

and taking $z = 1$ we have proved the lemma with $c_1 = 1 + \frac{2}{3}\lambda$. However, we exploit our inequality to the full when we choose c_1 as the minimum of the elementary function at the right side of the last inequality, which is given by

$$(46) \quad c_1^3 = 9\lambda/2.$$

One could also argue from the equation

$$(p\omega')' = 2(2\lambda - 1)f'p\omega.$$

If $\lambda \leq \frac{1}{2}$ then

$$(p\omega')' \leq 0, \quad p\omega' \leq -l,$$

and since $p \leq e^{iz^3}$,

$$\omega' \leq -l \cdot e^{-iz^3}, \quad 0 \leq \omega \leq 1 - l \cdot \int_0^z e^{-it^3} d\xi,$$

therefore

$$l \leq 1/B = 0.777 \quad \text{with} \quad B = \int_0^\infty e^{-iz^3} dz.$$

If, however, $\lambda \geq \frac{1}{2}$, then $f' \leq z$, $p\omega \leq 1$ yield

$$(p\omega')' \leq 2(2\lambda - 1)z, \quad p\omega' \leq -l + (2\lambda - 1)z^2$$

and taking the value

$$\int_0^\infty z^2 \cdot e^{-iz^3} dz = \int_0^\infty e^{-it^3} \cdot d(\frac{1}{3}z^3) = 1$$

into account, we obtain by the same argument

$$l \leq 2\lambda/B.$$

Hence the lemma is satisfied with

$$c_1 = 1/B \text{ for } \lambda \leq \frac{1}{2}, \quad c_1 = 2\lambda/B \text{ for } \lambda \geq \frac{1}{2}.$$

For small λ and large λ our first appraisal (46) is better, but in a certain middle range, namely for $0.104 \leq \lambda \leq 1.096$, the second gives a sharper result.

LEMMA 2.

$$(47) \quad 0 \leq -p\omega' \leq c_1 + c_2 z^2,$$

and thus, a fortiori,

$$(48) \quad 0 \leq -\omega' \leq c_1 + c_2 z^2.$$

PROOF.

$$(49) \quad p\omega' = \omega_1 - 2pf\omega \leq \omega_1 \leq 0.$$

$$(-p\omega')' = 2(1 - 2\lambda)f'p\omega.$$

Again distinguish the cases $\lambda \geq \frac{1}{2}$, $\lambda \leq \frac{1}{2}$. In the first case $-p\omega'$ decreases and therefore

$$-p\omega' \leq l \leq c_1.$$

In the second case

$$(-p\omega')' \leq 2(1 - 2\lambda)z,$$

$$-p\omega' \leq l + (1 - 2\lambda)z^2 \leq c_1 + (1 - 2\lambda)z^2.$$

Thus we have established Lemma 2 with

$$c_2 = 0 \text{ for } \lambda \geq \frac{1}{2}, \quad c_2 = 1 - 2\lambda \text{ for } \lambda \leq \frac{1}{2}.$$

The following two lemmas prepare the way for an asymptotic appraisal of $g(z)$ when z approaches infinity.

LEMMA 3.

$$(50) \quad \int_0^1 \omega(z) dz \geq c (> 0).$$

PROOF. Twice integrating (48) we find the inequalities

$$\omega(z) \geq 1 - c_1 z - \frac{1}{3}c_2 z^3,$$

$$\int_0^z \omega(\xi) d\xi \geq z - \frac{1}{2}c_1 z^2 - \frac{1}{12}c_2 z^4 \geq z(1 - \frac{1}{2}z/c_3)$$

for $z \leq 1$ where

$$1/c_3 = \max. (1, c_1 + \frac{1}{6}c_2).$$

Thus

$$\int_0^1 \omega(\xi) d\xi \geq \int_0^{c_3} \omega(\xi) d\xi = \frac{1}{2}c_3.$$

This argument is fully exploited by choosing $z = c_0$ as the point where the polynomial $1 - c_1 z - \frac{1}{3}c_2 z^3$ changes sign and then computing the area

$$\begin{aligned} c &= \int_0^{c_0} (1 - c_1 z - \frac{1}{3}c_2 z^3) dz = c_0(1 - \frac{1}{2}c_1 c_0 - \frac{1}{12}c_2 c_0^3) \\ &= \frac{1}{4}c_0(3 - c_0 c_1) (\geq \frac{1}{2}c_0), \end{aligned}$$

with the result

$$\int_0^{c_0} \omega(z) dz \geq c.$$

LEMMA 4. If $\int_0^1 g(z) dz \geq \gamma$ then

$$\left. \begin{aligned} 0 < \omega(z) &\leq e^{-\gamma(z-1)^2}, \\ 0 \leq -(\omega' + 2f\omega) &\leq (c_1 + c_2 z^2) \cdot e^{-\gamma(z-1)^2} \end{aligned} \right\} \text{ for } z \geq 1.$$

PROOF. The hypothesis implies for $z \geq 1$:

$$\begin{aligned} f'(z) &\geq \gamma, & f(z) &\geq \gamma(z-1), & 2 \int_0^z f(\xi) d\xi &\geq \gamma(z-1)^2, \\ p(z) &\geq e^{\gamma(z-1)^2}. \end{aligned}$$

Combine this with (45), (47) and (49): $-\omega_1 \leq -p\omega'$.

We topologize the functional space \mathcal{F} consisting of all functions $g = g(z)$ defined and continuous for $z \geq 0$ by agreeing that a sequence g_n approaches zero, $g_n \rightarrow 0$ with $n \rightarrow \infty$, if $g_n(z)$ converges to zero *uniformly in each finite interval*; in other words, if to every $\epsilon > 0$, $z > 0$ one can assign an $N(\epsilon, z)$ such that $0 \leq z \leq z_0$, $n \geq N(\epsilon, z_0)$ imply $|g_n(z)| \leq \epsilon$. This functional space \mathcal{F} is *complete*, i.e. convergence of a sequence g_n , $g_n - g_m \rightarrow 0$ with $n, m \rightarrow \infty$, implies convergence to some element g , $g_n - g \rightarrow 0$. Our domain \mathcal{G} defined by $0 \leq g(z) \leq 1$ is a closed convex part of \mathcal{F} . (See Appendix, under 1.)

Let $\delta(\epsilon, z)$ be any positive function of the variables $\epsilon > 0$, $z > 0$. The element $g \in \mathcal{G}$ is said to lie in \mathcal{G}_δ if the inequality $|g(z_1) - g(z_2)| \leq \epsilon$ holds whenever

$$0 \leq z_1, z_2 \leq z_0 \quad \text{and} \quad |z_1 - z_2| \leq \delta(\epsilon, z_0)$$

(equi-continuity of type δ). Clearly \mathcal{G}_δ is a *compact* subset of \mathcal{G} ; one has only to consider the values of g for rational arguments, marching them off in Indian file. The same simple argument of interpolation as employed by Birkhoff and Kellogg, l.c.,¹⁰ in proving their Theorem II (p. 103) yields the following general principle (see Appendix, under 2):

An operator $\Phi\{g\}$ defined and continuous in \mathcal{G} and mapping \mathcal{G} into \mathcal{G} , necessarily has a fixed element $g = \Phi\{g\}$.

According to (48), Lemma 2, our operator Φ_λ maps \mathcal{G} into the subset \mathcal{G}_δ corresponding to the function

$$\delta(\epsilon, z) = \epsilon / (c_1 + c_2 z^2).$$

Hence the existence of a solution g of the functional equation $g = \Phi_\lambda\{g\}$ will be proved as soon as we can establish continuity of the operator Φ_λ in \mathcal{G} :

LEMMA 5. *The image $\omega = \Phi_\lambda\{g\}$ depends continuously on $g \in \mathcal{G}$.*

This fact, which we are now going to prove, is less trivial than it appears, because it implies that ω varies but little when the "tail" of the function $g \in \mathcal{G}$ (i.e. its values for large values of the argument z) changes arbitrarily. The explicit formula (42) clearly shows that the particular solutions η , η_1 and ϑ , ϑ_1 depend continuously on g ; these functions in a finite interval $0 \leq z \leq z_0$ do not

depend on what $g(z)$ does for $z > z_0$. The salient point is the way in which the constant $l = -\omega'(0)$ in

$$\omega(z) = \eta(z) - l \cdot \vartheta(z)$$

depends on g .

Let $g^{(n)} \in \mathfrak{G}$ be a sequence which in the sense of our topology tends to g . Using the notations $\eta^{(n)}$, $\vartheta^{(n)}$, $l^{(n)}$, $\omega^{(n)}$ in an obvious way we have

$$\eta^{(n)} \rightarrow \eta, \quad \eta_1^{(n)} \rightarrow \eta_1; \quad \vartheta^{(n)} \rightarrow \vartheta, \quad \vartheta_1^{(n)} \rightarrow \vartheta_1 \quad \text{with } n \rightarrow \infty.$$

According to Lemma 1, all $l^{(n)}$ lie between 0 and c_1 and thus this sequence has at least one point of condensation l^* . The functions

$$\omega^*(z) = \eta(z) - l^* \cdot \vartheta(z), \quad \omega_1^*(z) = \eta_1(z) - l^* \cdot \vartheta_1(z),$$

being the limits of a subsequence of $\omega^{(n)}$, $\omega_1^{(n)}$, satisfy the inequalities $\omega^* \geq 0$, $\omega_1^* \leq 0$. However, we know that ω is the only solution of this kind, and consequently $l^* = l$. Thus the bounded sequence $l^{(n)}$ has only one condensation point, namely l , and therefore converges to l .

Our proof for the existence of a function $g \in \mathfrak{G}$ which is its own image under the operator Φ_λ is now complete. As an image it satisfies the inequalities (49), (50),

$$(51) \quad g' \leq 0, \quad \int_0^1 g(z) dz \geq c,$$

besides $0 \leq g \leq 1$, and as image of a g for which (51) holds, it satisfies the further conditions

$$(52) \quad g(z) \leq e^{-c(s-1)^2}, \\ 0 \leq -(g' + 2fg) \leq (c_1 + c_2 z^2) \cdot e^{-c(s-1)^2}$$

for $z \geq 1$ (Lemma 4). Expressing everything in terms of f we see that $f(0) = f'(0) = 0$, $f''(0) = 1$ and

$$(53) \quad f''' + 2ff'' - 2\lambda f'^2 = \text{const.}$$

Moreover f'' is monotone decreasing and f'' as well as $f''' + 2ff''$ tend to zero with $z \rightarrow \infty$ essentially as strongly as e^{-cz^2} . Hence the positive integral

$$\int_0^\infty g(z) dz = f'(\infty) = \beta$$

converges and for the constant on the right side of (53) we find the value $-2\lambda\beta^2$ so that

$$f''' + 2ff'' + 2\lambda(\beta^2 - f'^2) = 0.$$

An explicit appraisal of β is obtained from (51) and (52):

$$c < \beta < 1 + \sqrt{(\pi/2c)}.$$

Putting finally

$$w(z) = \kappa \cdot f(\kappa z) \quad \text{with} \quad \kappa = (k/\beta)^{\frac{1}{2}}$$

we may formulate our chief result as follows:

THEOREM. *For given positive k the problem (A_λ) has a solution w whose derivative is monotone increasing from 0 to k as z travels from 0 to infinity; the second derivative decreases monotonely from*

$$w''(0) = \alpha k^{\frac{1}{2}} \quad (\alpha = \beta^{-\frac{1}{2}})$$

to zero, approaching zero with $z \rightarrow \infty$ at least as strongly as a function of the type $e^{-\gamma z^{\frac{1}{2}}}$ ($\gamma > 0$).

So far so good. But I should like to see the aircraft engineer who will apply this method to compute the boundary layer for a given profile of an aerofoil!

We have *not* proved that there is only one solution of the problem (A_λ) . One could try to approach the question of uniqueness by studying the continuous variation of the operator Φ_λ with λ .¹¹ It seems not impossible to attack the general boundary layer equation (B_ν) by the method here developed.

8. Appendix

1. The *bounded* continuous functions g form a normed linear space \mathcal{F}_0 in which the norm $\|g\|$ induces our topology if, somewhat artificially, we define the norm by

$$\|g\| = \sum \frac{1}{2^\nu} \left\{ \max_{\nu-1 \leq z \leq \nu} |g(z)| \right\} \quad (\nu = 1, 2, \dots).$$

Whereas \mathcal{F}_0 is incomplete, the part \mathcal{G} is a complete closed subset of \mathcal{F}_0 . Hence the "general principle" on which we base our argument will fit into Schauder's scheme only if one slightly generalizes his central theorem (Satz 2, on p. 175 of *Studia Mathematica* 2, 1930) in the following manner. Let \mathcal{F}_0 be a normed linear space; a continuous mapping of a complete and closed convex subset \mathcal{G} of \mathcal{F}_0 into a compact subset \mathcal{G}^* of \mathcal{G} has a fixed point.

2. In adapting the proof to our conditions, I give it a more constructive twist. For the open interval $0 \leq z < \infty$ one has to combine ever more refined subdivision with exhaustion. Let therefore n, ν be two positive integers. For any given numbers

$$x_m \quad (m = 0, 1, \dots, n\nu; 0 \leq x_m \leq 1)$$

form by *linear interpolation* the function $g(z) = g(z; x_0, \dots, x_{n\nu})$ with the prescribed values

$$g(m/n) = x_m \quad (m = 0, 1, \dots, n\nu)$$

in the interval $0 \leq z \leq \nu$ and extrapolate it beyond ν by $g(z) = x_{n\nu}$ for $z \geq \nu$. Denote by x_m^* the values of its image $g^* = \Phi\{g\}$ at the points $z = m/n$ ($m =$

¹¹ See E. Rothe, *Bull. Am. Math. Soc.* 45, 1939, pp. 606-613.

$0, 1, \dots, n\nu$). The continuous mapping $x_m \rightarrow x_m^*$ of the $(n\nu + 1)$ -dimensional unit cube $0 \leq x_0 \leq 1, \dots, 0 \leq x_{n\nu} \leq 1$ has a fixed point (Brouwer's Theorem); choose one, let it have the coordinates x_m^0 and set

$$g_{n,\nu}(z) = g(z; x_0^0, \dots, x_{n\nu}^0).$$

The image $g_{n,\nu}^*$ of $g_{n,\nu}$ takes on the same values x_m^0 as $g_{n,\nu}$ itself at the points $z = m/n$. Let ν_0 be a positive integer. Since $g_{n,\nu}^* \in \mathfrak{G}_\delta$,

$$(54) \quad |g_{n,\nu}^*(z) - x_m^0| \leq \epsilon \quad \text{for} \quad \frac{m}{n} \leq z \leq \frac{m+1}{n} \quad (m = 0, 1, \dots, n\nu_0 - 1)$$

provided $\nu \geq \nu_0$ and $1/n \leq \delta(\epsilon, \nu_0)$; in particular $\left(z = \frac{m+1}{n}\right)$

$$|x_{m+1}^0 - x_m^0| \leq \epsilon.$$

Hence, because $g_{n,\nu}(z)$ is the linear interpolation of the values x_m^0 , the inequality

$$|g_{n,\nu}(z) - x_m^0| \leq \epsilon$$

holds under the same conditions as (54) and thus

$$|g_{n,\nu}^*(z) - g_{n,\nu}(z)| \leq 2\epsilon \quad \text{for } 0 \leq z \leq \nu_0$$

as soon as $\nu \geq \nu_0$ and $n \geq 1/\delta(\epsilon, \nu_0)$. This means that $g_{n,\nu}^* - g_{n,\nu} \rightarrow 0$ with n and ν tending to infinity. A subsequence¹² of the $g_{n,\nu}^* \in \mathfrak{G}_\delta$ tends to a limit g , the corresponding subsequence of the $g_{n,\nu}$ to the same limit, and, on account of the continuity of Φ , the relation $g_{n,\nu}^* = \Phi\{g_{n,\nu}\}$ yields $g = \Phi\{g\}$.

INSTITUTE FOR ADVANCED STUDY

¹²By a subsequence of the pairs (n, ν) we mean a sequence (n_i, ν_i) both members of which are monotone: $n_i < n_{i+1}$, $\nu_i < \nu_{i+1}$.

ABSOLUTELY CONVERGENT FOURIER EXPANSIONS FOR NON-COMMUTATIVE NORMED RINGS

BY S. BOCHNER AND R. S. PHILLIPS

(Received February 11, 1942)

I. The main theorem

The set R of elements x, y, \dots will be called a *normed ring* if

- a) R is a Banach space over the field of complex numbers.
- b) The operation of multiplication is defined for elements of R . This operation possesses the usual algebraic properties.
- c) R contains a unit element e .
- d) $\|xy\| \leq \|x\| \cdot \|y\|$, $\|e\| = 1$.

Since commutativity of multiplication is not assumed we shall have to distinguish between left inverses, right inverses, and (two-sided) inverses. We note that a left inverse, if it exists, need not be unique.

Our principal result is the following theorem:

THEOREM 1. *If R' is the ring of periodic functions*

$$(1) \quad x(t) = \sum_{-\infty}^{\infty} a_n e^{int}, \quad a_n \in R,$$

on $0 \leq t < 2\pi$ to R , with

$$(2) \quad \sum \|a_n\| < \infty,$$

where the product $x(\cdot) \cdot y(\cdot)$ in R' is the function $x(t)y(t)$, then $x(t)$ has a left inverse in R' if $x(t_0)$ has a left inverse in R , for every t_0 .

This is a generalization of the known theorem of N. Wiener¹ from complex numbers to a general R .

Denoting by $L(R)$ the Banach space of strongly integrable [Bochner, 2] functions $f(t)$ on $-\infty < t < \infty$ to R with the norm

$$(3) \quad \|f(\cdot)\| = \int_{-\infty}^{\infty} \|f(t)\| dt$$

we are able to obtain a generalization of another of Wiener's theorems.

THEOREM 2. *If $f(t) \in L(R)$ and if the Fourier transform*

$$(4) \quad x(u) = \int_{-\infty}^{\infty} f(t) e^{iut} dt$$

has a left inverse in R , for each real u , then the linear combinations

$$(5) \quad \sum_n A_n f(t - \lambda_n), \quad A_n \in R,$$

¹ We shall quote from the memoir of Wiener [6] rather than from his book on the Fourier Integral.

are dense in $L(R)$. Conversely, if (5) is dense in $L(R)$ then $x(u)$ has a left inverse throughout $(-\infty, \infty)$.

The possibility of extending Wiener's results from numbers to ring elements was suggested to us by Gelfand's new method [3] of proving Wiener's theorem with the aid of maximal ideals in rings. It turns out that Wiener's own method leads very directly to our version of his theorem, whereas the adaptation of Gelfand's method to the non commutative case is slightly more elaborate. We shall present both methods. The normed ring approach has been found applicable to a slightly more general type of function ring.

II. Wiener's method

LEMMA 1. If $a \in R$, $\|a\| < 1$, then the series

$$c = e - a + a^2 - a^3 + \dots$$

is absolutely convergent and

$$(6) \quad c(e + a) = (e + a)c = e.$$

Relation (6) can be verified by direct multiplication.

LEMMA 2. If a has a left inverse a' (i.e. $a'a = e$) and if $\|b\| \cdot \|a'\| < 1$, then $a + b$ has a left inverse c , and

$$(7) \quad c = a'(e - ba' + (ba')^2 - (ba')^3 + \dots)$$

In fact, $a + b = (e + ba')a$, and hence, by Lemma 1,

$$c(a + b) = a'(e - ba' + (ba')^2 - \dots)(e + ba')a = a'ea = e.$$

LEMMA 3. The set U_l of all elements in R with a left inverse is an open set.

PROOF. Lemma 2.

Since R is a Banach space, we first note that if $x(t)$ belongs to the R' of Theorem 1, it is in particular strongly integrable. Therefore series (1) is its Fourier series, and many familiar theorems hold. For instance, expansion (1) is unique, no matter how arrived at. In what follows, analysis of real and complex variables will be applied to the value space R without special emphasis.

LEMMA 4. If $x(\cdot) \in R'$, and if $x(0)$ has a left inverse in R , then there exists an element

$$y(t) = \sum c_n e^{int}$$

in R' with the following two properties:

(i) the coefficient c_0 has a left inverse c'_0 , and

$$\|c'_0\| \cdot \left\| \sum_{n=1}^{\infty} (c_n + c_{-n}) \right\| < 1;$$

(ii) in some interval $-\epsilon < t < \epsilon$, $y(t) = x(t)$.

PROOF. As in Wiener [6, p. 12, Lemma 2d], we introduced on the circle $-\pi \leq t < \pi$ the numerical function

$$\omega_\epsilon(t) = \begin{cases} 1, & |t| < \epsilon \\ 2 - \frac{|t|}{\epsilon}, & \epsilon \leq |t| < 2\epsilon \\ 0, & 2\epsilon \leq |t|; \end{cases}$$

and the function

$$y_\epsilon(t) = \omega_\epsilon(t) \cdot x(t) + [1 - \omega_\epsilon(t)] \cdot x(0) = \sum c_n(\epsilon) e^{int}.$$

Obviously $y_\epsilon(t)$ satisfies property (ii) for each ϵ . As for property (i), Wiener's argument shows that

$$\lim_{\epsilon \rightarrow 0} c_0(\epsilon) = x(0),$$

$$\lim_{\epsilon \rightarrow 0} \sum_{n=1}^{\infty} \|c_n(\epsilon) + c_{-n}(\epsilon)\| = 0.$$

Thus by Lemma 2 property (i) holds for sufficiently small ϵ .

LEMMA 5. If $y(\cdot) \in R'$, and $y(t)$ satisfies property (i) of Lemma 4, then $y(\cdot)$ has a left inverse $y'(\cdot)$ in R' .

In fact, putting

$$B(t) = \sum_{n=1}^{\infty} (c_n e^{int} + c_{-n} e^{-int}),$$

we have by Lemma 2,

$$y'(t) = c'_0[e - (B(t)c'_0 + (B(t)c'_0)^2 - (B(t)c'_0)^3 + \dots]$$

Now if $x(\cdot) \in R'$ and if for all t , $x(t) \in U_l$, then by Lemmas 4 and 5 there exists for each t_0 a function $y_{t_0}(\cdot) \in R'$ such that $y_{t_0}(t) \cdot x(t) = e$ in some interval $(t_0 - \epsilon, t_0 + \epsilon)$. We can now piece together a finite number of these functions $y_{t_0}(\cdot)$ and obtain a function $x'(\cdot) \in R'$ for which $x'(t) \cdot x(t) = e$ for all t . For details of the argument see Wiener [6, pp. 10–11, Lemma II_b]. This concludes the proof of Theorem 1.

We now turn to the proof of Theorem 2. Our main step is

LEMMA 6. Given constants $-\pi < \alpha < a < b < \beta < \pi$, if $x_1(\cdot)$, $x_2(\cdot)$ belong to R' where $x_2(u)$ has a left inverse for $\alpha < u < \beta$, and $x_1(u)$ vanishes outside $a \leq u \leq b$, then there exists an element $x_3(\cdot)$ of R' which vanishes outside $\alpha \leq u \leq \beta$ such that

$$(8) \quad x_1(u) = x_3(u)x_2(u)$$

in $-\pi \leq u < \pi$.

PROOF. By Lemmas 4 and 5 corresponding to any u_0 in $a \leq u \leq b$, there exists a function $y_{u_0}(\cdot) \in R'$ such that in some interval $(u_0 - \epsilon, u_0 + \epsilon)$, $y_{u_0}(u)x_2(u) = e$. As above, we can define a function $x'_2(u)$ which belongs to R'

and for which $x'_2(u)x_2(u) = e$ in $a \leq u \leq b$. Let $x_3(u) = x_1(u) \cdot x'_2(u)$. Then $x_3(u) = 0$ in the intervals $-\pi \leq u \leq \alpha$, $\beta \leq u < \pi$. In addition

$$x_3(u) \in R', \quad \text{and} \quad x_1(u) = x_3(u)x_2(u).$$

LEMMA 7. If $x(u)$ is strongly integrable in $(-\pi, \pi)$ and vanishes in $(-\pi, -\pi + \epsilon)$ and $(\pi - \epsilon, \pi)$, and if we put

$$f(t) = \frac{1}{2\pi} \int_{-\pi}^{\pi} x(u)e^{-iut} du, \quad a_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} x(u)e^{-inu} du$$

then $\int_{-\infty}^{\infty} \|f(t)\| dt < \infty$ if and only if $\sum \|a_n\| < \infty$.

PROOF. As in Wiener [6, p. 14-15, Lemma IIf].

From Lemmas 6, 7 we now conclude

LEMMA 8. If $g(t)$ and $f(t)$ both belong to $L(R)$, if $x_1(u) = \int_{-\infty}^{\infty} g(t)e^{-iut} dt$ vanishes outside some interval (a, b) , and if $x_2(u) = \int_{-\infty}^{\infty} f(t)e^{-iut} dt$ vanishes outside some larger interval (α, β) , $\alpha < a$, $b < \beta$, $x_2(u)$ having a left inverse for each u in $\alpha < u < \beta$, then there exists an element $h(t)$ in $L(R)$ such that

$$(9) \quad g(t) = \int_{-\infty}^{\infty} h(\lambda)f(t - \lambda) d\lambda.$$

Integral (9) can be approximated in $L(R)$ by sums of the form (5). The transition from Lemma 8 to Theorem 2 can be effected by Fejer approximation as in Wiener [6, p. 16-18].

III. Irreducibility and maximal ideals

The following lemmas are the algebraic basis for an adaptation of Gelfand's methods.

Let $V = (a, b, c, \dots)$ be a Banach space over the complex numbers, and let $A = (\alpha, \beta, \dots)$ be the ring of all bounded linear transformations on V to V . Then $\alpha(a + b) = \alpha a + \beta b$, $(\beta\alpha)a = \beta(\alpha a)$. If V_0 is a subset of V and A_0 a subset of A then A_0V_0 will denote the set of all elements (αa) of V for $\alpha \in A_0$ and $a \in V_0$.

V will be said to be *irreducible* over the subring A_0 of A if for every $a \in V$, ($a \neq 0$), $A_0a = V$. This means that V has no subspaces invariant with respect to A_0 .

LEMMA 9. If V is irreducible over A_0 , and if A' denotes the set of all elements of A which commute with all elements of A_0 , then A' is isomorphic to the field of complex numbers.

In fact if $\alpha', \beta' \in A'$ and $\alpha_0 \in A_0$, then $\alpha'\alpha_0 = \alpha_0\alpha'$, $\beta'\alpha_0 = \alpha_0\beta'$ implies $(\alpha' \pm \beta')\alpha_0 = \alpha_0(\alpha' \pm \beta')$. Similarly $\alpha'\beta'\alpha_0 = \alpha'\alpha_0\beta' = \alpha_0\alpha'\beta'$. Thus A' is a ring. Clearly A' contains the null element and the identity. Now, if $\alpha' \in A'$, $\alpha' \neq 0$, consider the set $W = \alpha'V$. The irreducibility implies $A_0W = A_0V = V$,

and the commutativity implies $\alpha' A_0 = A_0 \alpha'$. Therefore $W = \alpha' V = \alpha'(A_0 V) = A_0(\alpha' V) = A_0 W = V$. That is, the range of the homomorphism α' is all of V . Now suppose that for some $a \in V$, $a \neq 0$, we had $\alpha' a = 0$. Then $\alpha' V = \alpha'(A_0 a) = A_0(\alpha' a) = 0$ which is contrary to $\alpha' V = V$. α' is therefore a one to one linear bounded transformation on V to V . By a theorem due to Banach (1, p. 41, Theorem 5), the inverse β' is a linear bounded transformation and hence belongs to A . $\beta' \alpha_0 = \beta' \alpha_0 \alpha' \beta' = \beta' \alpha' \alpha_0 \beta' = \alpha_0 \beta'$ for all $\alpha_0 \in A_0$. Thus β' belongs to A' . A' is therefore a field. Clearly a limit (in A) of transformations which commute with elements of A_0 also commutes with these elements. A' is thus a complete normed field over the complex numbers. Such a field is isomorphic to the field of complex numbers. Since the proof of this is the same in the commutative and noncommutative cases we refer the reader to Gelfand [3, pp. 6-8].

If R is a ring with unit element, a *left ideal* I is a subset with the properties:

1) if $x \in I$, $y \in I$, then $px + qy \in I$ for any elements p, q from R .

2) I is a proper subset of R .

A *maximal left ideal* is one not contained in a larger left ideal.

LEMMA 10. If R is a ring with unit, if I is a maximal left ideal, if V is the addition group of the cosets R/I , and if A_0 is the ring of homomorphisms of V onto itself as produced by multiplying V by elements of R from the left, then V is irreducible with respect to A_0 .

If $x \in R$, then the element of A_0 corresponding to x will be denoted by $\mu(x)$. For fixed $a \in V$, $a \neq 0$, we form the set $V_0 = \mu(R)a$. Since R is a linear space, V_0 is a linear subspace of V . Also, since R contains a unit element e , $\mu(e)a = a \neq 0$. Thus V_0 contains the null element and some other elements. Now form the set S of all elements of R contained in the cosets composing V_0 . Then S includes I as a proper part. On the other hand, S is a left ideal. As I was supposed maximal, $S = R$. Thus $V_0 = V$; that is $A_0 a = V$.

LEMMA 11. If R, I, V, A_0 have the same meaning as in Lemma 10, and if for fixed $x \in R$ and every maximal left ideal I the corresponding element $\mu(x)$ of A_0 has a left inverse in A_0 , then x has a left inverse in R .

PROOF. Since R contains a unit element e , every left ideal is contained in a maximal left ideal, and consequently an element x of R has a left inverse x' if (and only if) it is contained in no maximal ideal. See Gelfand [3, p. 8-9].

If $x'x = e$, then $\mu(x')\mu(x) = \mu(e)$. Denoting by a_e and a_0 the elements of V which as cosets in R/I contain the elements e and 0 respectively, we have in particular

$$(10) \quad \mu(x')\mu(x)a_e = \mu(e)a_e.$$

Since $e \cdot e = e$, we have $\mu(e)a_e = a_e$. On the other hand, $\mu(x)a_e$ contains x . Hence if x were contained in $a_0 = I$, we would have $\mu(x)a_e = a_0$, and consequently $\mu(x')\mu(x)a_e = a_0$. This completes the proof of Lemma 11.

Finally we have to make several statements for *normed rings*.

LEMMA 12. Every maximal ideal of a normed ring is closed.

For the proof see Gelfand [3, p. 8].

Preparatory to an amendment of Lemma 11, we first observe that if R is normed, and X is a closed linear subspace, then R/X is made into a Banach space by introducing the norm

$$\|Y\| = \inf_{y \in Y} \|y\|$$

for any Y of R/X [see 3].

LEMMA 13. If R, I, V, A_0 , have the same meaning as in Lemma 10, then V is a Banach space, A_0 is a normed ring, and for the norm $\|\mu(x)\|$ in A_0 we have

$$\|\mu(x)\| \leq \|x\|, \quad x \in R.$$

In fact,

$$\|\mu(x)\| = \sup_{a \in V} \frac{\|\mu(x)a\|}{\|a\|} = \sup_{a \in V} \frac{\inf_{y \in a} [\|xy\|]}{\inf_{y \in a} [\|y\|]} \leq \|x\|.$$

IV. Theorem 3

ASSUMPTIONS. Let F denote a (commutative) ring of complex valued functions $f(t)$ on a point set $[t]$. Ring multiplication is ordinary multiplication $f(t) \cdot g(t)$, and the constant function $f(t) \equiv 1$ belongs to F and is its unit. The norm in F will be denoted by ordinary bars: $|f(t)|$. By $M(f)$ we shall denote any continuous ring homomorphism from F to complex numbers. Thus in particular $M(fg) = M(f) \cdot M(g)$.

Let $R = (x, y, \dots)$ denote any (non-commutative) normed ring with unit e and norm $\|x\|$, and let R' denote a family of functions $x(\cdot) = x(t)$ from $[t]$ to R with the following properties:

- 1) R' is a ring under point multiplication $x(t)y(t)$.
- 2) If $x_1, \dots, x_n \in R, f_1(t), \dots, f_n(t) \in F$, then

$$(11) \quad x_1 f_1(\cdot) + \dots + x_n f_n(\cdot)$$

belongs to R' . If $x \in R, f(t) \in F$, then the element $xf(\cdot)$ of R' will be denoted by x' .

- 3) R' is a normed ring with norm $\|x(\cdot)\|$ for which

$$\|x'\| = \|x\| \cdot |f|.$$

- 4) The linear combinations (11) are dense in R' .

- 5) If $x(\cdot) = x'_1 + \dots + x'_n$, then for every homomorphism $M(f)$ of F ,

$$\|x_1 M(f_1) + \dots + x_n M(f_n)\| \leq \|x(\cdot)\|.$$

On the basis of 4) and 5) every $M(f)$ gives rise to a continuous homomorphism $\mathbf{M}(x(\cdot))$ from R' to R , with the property

$$\mathbf{M}(x') = xM(f).$$

We will call \mathbf{M} a generated homomorphism.

CONCLUSION. An element $x(\cdot) \in R'$ has a left inverse if for every generated homomorphism \mathbf{M} , the element $\mathbf{M}(x(\cdot))$ of R has a left inverse in R .

PROOF. Replacing R by R' in Lemma 11, we consider an arbitrary maximal left ideal I and for any element $x(\cdot) \in R'$ the corresponding element $\mu(x(\cdot))$ of A_0 . The elements $e'(\cdot) = ef(\cdot)$ of R' , $e = \text{unit in } R$, $f \in F$, are commutative with any element x'' of R' and hence by 4) with every element of R' . Therefore, by Lemmas 9, 10, and 13, $\mu(e'(\cdot))$ can be looked upon as a continuous ring homomorphism from F to complex numbers. More precisely there exists an $M(f)$ such that

$$\mu(e') = e_1 M(f) \quad (e_1 = \mu(e \cdot 1) \text{ is unit of } A).$$

For an element of the form (11) we have

$$\begin{aligned} \mu(x(\cdot)) &= \mu(x_1' + \cdots + x_k') = \mu(x_1 \cdot 1) \mu(e_{f_1}) + \cdots + \mu(x_k \cdot 1) \mu(e_{f_k}) \\ &= \mu(x_1 \cdot 1) M(f_1) + \cdots + \mu(x_k \cdot 1) M(f_k) = \mu([x_1 M(f_1) + \cdots + x_k M(f_k)] \cdot 1) \end{aligned}$$

and thus

$$(12) \quad \mu(x(\cdot)) = \mu(\mathbf{M}(x(\cdot)) \cdot 1)$$

By property 5) in conjunction with Lemma 13 this relation is valid for all elements $x(\cdot) \in R'$. Now if $\mathbf{M}(x(t))$ has a left inverse x' in R , then

$$\mu(x' \cdot 1) \cdot \mu(x(\cdot)) = \mu(x' \cdot x(\cdot)).$$

Applying (12) to $x' \cdot x(\cdot)$ instead of $x(\cdot)$, this is

$$= \mu(\mathbf{M}(x' \cdot (x(\cdot))) \cdot 1) = \mu(x' \cdot \mathbf{M}(x(\cdot)) \cdot 1) = \mu(e \cdot 1) = e_1.$$

Thus $\mu(x(\cdot))$ has a left inverse, and by Lemma 11, $x(\cdot)$ has a left inverse in R' .

A PARTICULAR ASSUMPTION. Corresponding to any $M(f)$ there exists a point t_0 such that

$$M(f) = f(t_0) \quad f \in F.$$

In this case, property 5) can be replaced by the simpler property: $\|x(t_0)\| \leq \|x(\cdot)\|$, for each t_0 .

A PARTICULAR CONCLUSION. The element $x(\cdot) \in R'$ has a left inverse in R' , provided $x(t)$ has a left inverse in R , for all t .

The proof is obvious.

Adjunction of unit. If the ring F has no unit, we make a formal adjunction of a unit 1, and we consider the enlarged ring $\bar{F}: \lambda 1 + f$, with the norm $|\lambda| + |f|$ ($\lambda = \text{complex number}$). This leads to the ring \bar{R}' of elements $\bar{x}(\cdot) = \lambda e \cdot 1 + x(\cdot)$ where $\|\bar{x}(\cdot)\| = |\lambda| + \|x(\cdot)\|$. It can be shown, Gelfand [3], that every homomorphism $\bar{M} = \bar{M}(\lambda 1 + f)$ to the complex numbers is either $\lambda + M(f)$ where M is a homomorphism of F , or it is the exceptional homomorphism $\bar{M}(\lambda 1 + f) = \lambda$. We can then obtain the following conclusion: If F has no unit, then the element $\bar{x}(\cdot) = \lambda e \cdot 1 + x(\cdot)$ with $\lambda \neq 0$ and $x(\cdot) \in R'$, has a left inverse of the form $\lambda' e \cdot 1 + x'(\cdot)$, if for every generated homomorphism \mathbf{M} the element $\lambda e + \mathbf{M}(x(\cdot)) \in R$ has a left inverse in R . In the particular circumstances cited before, it is again sufficient that $\lambda e + x(t)$ shall have an inverse for each t .

V. Expansions on groups. Applications

By known results on topological groups, and by results on numerical functions by Gelfand [4], and Gelfand-Rykov [5], Theorem 3 leads easily to the following theorems.

THEOREM 4. *If $\Gamma = (\alpha, \beta, \dots)$ is a commutative group of addition with discrete topology and $G = (t, s, \dots)$ is the compact dual group of addition; if $\{\chi(\alpha, t)\}$, $\alpha \in \Gamma$, $t \in G$ are the characters from Γ to G ; if R is any normed ring; and if R' is the ring of functions on G ,*

$$(13) \quad x(t) = \sum_{\alpha} a_{\alpha} \chi(\alpha, t) \quad a_{\alpha} \in R$$

$$(14) \quad \sum_{\alpha} \|a_{\alpha}\| < \infty,$$

then $x(\cdot)$ has a left inverse in R' , provided $x(t)$ has a left inverse in R , for all t .

More generally, if (14) is replaced by

$$\sum_{\alpha} e^{q_{\alpha}} \|a_{\alpha}\| < \infty,$$

where the real numbers q_{α} have the properties $q_{\alpha+\beta} \leq q_{\alpha} + q_{\beta}$ and $q_0 = 0$, then the condition is that

$$\sum_{\alpha} a_{\alpha} e^{p_{\alpha}} \chi(\alpha, t)$$

shall have a left inverse in R for all t , and every system of real numbers p_{α} , for which $p_{\alpha+\beta} = p_{\alpha} + p_{\beta}$, $p_0 = 0$, $p_{\alpha} \leq q_{\alpha}$.

However, if Γ is a locally bicomact group with unique Haar measure $d\alpha$; if $\chi(\alpha, t)$ are continuous characters; if R' is formed by

$$x(t) = \int_{\Gamma} a_{\alpha} \chi(\alpha, t) d\alpha, \quad a_{\alpha} \in R$$

where

$$\int_{\Gamma} e^{q_{\alpha}} \|a_{\alpha}\| d\alpha < \infty;$$

and if R' has no unit, but 1 is an adjoined unit, then the element $\bar{x}(\cdot) = \lambda 1 + x(\cdot)$ has a left inverse of the form $\lambda' 1 + x'(\cdot)$ provided

$$\lambda \cdot \left[\lambda \cdot e + \int_{\Gamma} a_{\alpha} e^{q_{\alpha}} \chi(\alpha, t) d\alpha \right]$$

has a left inverse for all t , and every continuous system p_{α} with the previous properties.

In the case of a discrete group Γ it is often natural to consider not the total compact group G but a dense subgroup G_0 . For instance, if Γ is the linear group $-\infty < \alpha < \infty$ without topology then it is natural to consider the characters $\chi(\alpha, t) = e^{i\alpha t}$ with $-\infty < t < \infty$, dense in the total group G . It is not

necessary to make sure of the existence of $x'(t)$ for all G , but it is sufficient to know that the left inverse $x'(t)$ exists and is bounded in norm $(-\infty, \infty)$. This results from the following lemma:

LEMMA 14. Suppose $x_n \rightarrow x$ in R . If x'_n is a left inverse of x_n and the $\|x'_n\|$ are bounded, then x possesses a left inverse.

PROOF. $e - x'_n x = e - x'_n x_n + x'_n x_n - x'_n x$. Hence $\|e - x'_n x\| \leq \|x'_n\| \cdot \|x_n - x\| \rightarrow 0$. By Lemma 1 there exists for sufficiently large n , an inverse y_n of $x'_n x$ so that $y_n y'_n$ is a left inverse for x .

However, if Γ is the continuous group $-\infty < \alpha < \infty$ with the ordinary Lebesgue measure as measure, the line $(-\infty, \infty)$ is the complete group G . Thus we obtain the following:

THEOREM 5. If $x(t) = \sum a_n e^{i\alpha_n t}$, $a_n \in R$, $\sum \|a_n\| < \infty$, and if a left inverse $y(t)$ exists for each t and $\|y(t)\| < C$, then there exists a function $x'(t) = \sum b_n e^{i\beta_n t}$ with $\sum \|b_n\| < \infty$ such that $x'(t) \cdot x(t) = e$.

If $a(\alpha) \in R$, $-\infty < \alpha < \infty$, and $a(\alpha)$ is strongly integrable $\int_{-\infty}^{\infty} \|a(\alpha)\| d\alpha < \infty$, and if for complex $\lambda \neq 0$, $\lambda e + \int_{-\infty}^{\infty} a(\alpha) e^{i\alpha t} d\alpha$ has a left inverse for all t , then there exists a left inverse of the form $\lambda' e + \int_{-\infty}^{\infty} b(\alpha) e^{i\alpha t} d\alpha$, with $\int_{-\infty}^{\infty} \|b(\alpha)\| d\alpha < \infty$.

Difference and integral equations. Now, consider functions $\varphi(t)$, $\psi(t)$, ... on the line $-\infty < t < \infty$ with values in a space on which R operates; and assume that they form a Banach space B , and that the norm in B is invariant under translation. Then, all symbols having the same meaning as in Theorem 5, we see that the difference equation

$$(15) \quad \sum a_n \varphi(t - \alpha_n) = \psi(t)$$

has a solution

$$(16) \quad \varphi(t) = \sum b_n \psi(t - \beta_n),$$

and the integral equation

$$\varphi(t) + \int_{-\infty}^{\infty} a(\alpha) \varphi(t - \alpha) d\alpha = \psi(t)$$

has a solution

$$\varphi(t) = \psi(t) + \int_{-\infty}^{\infty} b(\beta) \psi(t - \beta) d\beta.$$

For instance, let R be the space of k -dimensional matrices of complex numbers

$$x \equiv (x_{ij}) \quad i, j = 1, \dots, k,$$

and let $\|x\|$ be the maximum of $\sum_{i,j=1}^k x_{ij} p_i q_j$ for $\sum |p_i|^2 \leq 1$, $\sum |q_j|^2 \leq 1$. Then we obtain the following result.

If $a_n = (a_{ij}^n)$, $\sum \|a_n\| < \infty$, $\alpha_n = \text{real}$, and if for

$$D(t) = \left(\sum_{n=1}^{\infty} a_{ij}^n e^{i\alpha_n t} \right) \quad i, j = 1, \dots, k,$$

$$\|D^{-1}(t)\| \leq C < \infty \quad -\infty < t < \infty,$$

then there exist matrices $b_n = (b_{ij}^n)$, $\sum \|b_n\| < \infty$, and real numbers β_n , such that the system of equations

$$\sum_{n=1}^{\infty} \sum_{j=1}^k a_{ij}^n \varphi_j(t + \alpha_n) = \psi_i(t) \quad i = 1, \dots, k$$

has a solution of the form

$$\varphi_i(t) = \sum_{n=1}^{\infty} \sum_{j=1}^k b_{ij}^n \psi_j(t + \beta_n).$$

If the functions $\psi_j(t)$ belong to Lebesgue class L_p , to B , M , C etc., then the functions $\varphi_i(t)$ belong to the same class respectively.

PRINCETON UNIVERSITY,
HARVARD UNIVERSITY

REFERENCES

1. S. BANACH, *Théorie des Operations Linéaires*, Warsaw, 1932.
2. S. BOCHNER, *Integration von Funktionen, deren Werte die Elemente eines Vektorraumes sind*, Fund. Math., vol. 20, 1933, pp. 262-276.
3. I. GELFAND, *Normierte Ringe*, Recueil Mathématique, vol. 9 (51), 1941, pp. 3-24.
4. I. GELFAND, *Über absolut konvergente trigonometrische Reihen und Integrale*, Recueil Mathématique, vol. 9 (51), 1941, pp. 51-66.
5. I. GELFAND AND D. RAIKOV, *On the Theory of Characters of Commutative Topological Groups*, Comptes Rendus de l'Académie des Sciences de l'URSS, vol. 28, 1940, pp. 195-6.
6. N. WIENER, *Tauberian theorems*, these Annals, vol. 33, 1932, pp. 1-100.

ON THE LAW OF THE ITERATED LOGARITHM

BY PAUL ERDÖS

(Received December 26, 1941)

Introduction

Let t be a real number ($0 \leq t \leq 1$), and let $t = 0.\epsilon_1(t)\epsilon_2(t) \cdots$ be its dyadic expansion, or equivalently,

$$(0.1) \quad t = \frac{\epsilon_1(t)}{2} + \frac{\epsilon_2(t)}{2^2} + \cdots + \frac{\epsilon_n(t)}{2^n} + \cdots,$$

where $\epsilon_n(t) = 0$ or 1 according as the integral part of $2^n t$ is even or odd. It is well known that $\{\epsilon_n(t)\}$ ($n = 1, 2, \cdots$) is an independent system in the sense of probability,¹ and that

$$(0.2) \quad \int_0^1 \epsilon_n(t) dt = \frac{1}{2}, \quad \int_0^1 \left(\epsilon_n(t) - \frac{1}{2} \right)^2 dt = \frac{1}{4}.$$

Let us further put

$$(0.3) \quad f_n(t) = \sum_{k=1}^n \epsilon_k(t) - \frac{n}{2}.$$

It was proved by A. Khintchine² and A. Kolmogoroff³ that

$$(0.4) \quad \limsup_{n \rightarrow \infty} \frac{f_n(t)}{\left(\frac{n}{2} \log \log n \right)^{\frac{1}{2}}} = 1$$

for almost all t .

Let $\varphi(n)$ be a monotone increasing non-negative function defined for all sufficiently large integers. Following P. Lévy we say that $\varphi(n)$ belongs to the *upper class* if, for almost all t , there exist only finitely many n such that

$$(0.5) \quad f_n(t) > \varphi(n);$$

and $\varphi(n)$ belongs to the *lower class* if, for almost all t , there exist infinitely many n such that (0.5) is true. According to the well-known law of 0 or 1, each $\varphi(n)$ must belong to one of these classes. Then the result of A. Khintchine and A. Kolmogoroff stated above means that $\varphi(n) = (1 + \epsilon)(\frac{1}{2}n \log \log n)^{\frac{1}{2}}$ belongs to the upper class if $\epsilon > 0$, and to the lower class if $\epsilon < 0$.

The purpose of the present paper is to give a sharpening of this result. The

¹ Cf. M. Kac and H. Steinhaus, *Sur les fonctions indépendentes*, Studia Math. 6 (1936), 46-58, 59-66, 89-97.

² A. Khintchine, *Asymptotische Gesetz der Wahrscheinlichkeitsrechnung*, Berlin, 1933.

³ A. Kolmogoroff, *Über das Gesetz der iterierten Logarithmus*, Math. Annalen, 101 (1929), 126-135.

main results are stated in Theorems 1, 2, 3, 4, and 5 below. Among other results, it follows from Theorem 3 that, for $k > 3$,

$$(0.6) \quad \varphi(n) = \left(\frac{n}{2 \log \log n} \right)^{\frac{1}{2}} \left(\log \log n + \frac{3}{4} \log_3 n + \frac{1}{2} \log_4 n \right. \\ \left. + \cdots + \frac{1}{2} \log_{k-1} n + \left(\frac{1}{2} + \epsilon \right) \log_k n \right)$$

belongs to the upper class if $\epsilon > 0$ and to the lower class if $\epsilon \leq 0$.

Our proof is direct and elementary. We do not assume the result of A. Khintchine and A. Kolmogoroff, and the paper can be read without knowledge of any particular results concerning the law of the iterated logarithm. The only facts we need are the notion of independence, and the well known inequality

$$(0.7) \quad c_1 \frac{n}{x} e^{-2x^2/n} < Pr(A_n(x)) < c_2 \frac{n}{x} e^{-2x^2/n},$$

where

$$(0.8) \quad A_n(x) = E[t: f_n(t) > x]$$

means the set of all real numbers t ($0 \leq t \leq 1$) satisfying $f_n(t) > x$, and $Pr(A)$ means the ordinary Lebesgue measure of a measurable set A in the interval $0 \leq t \leq 1$. c_i ($i = 1, 2, \dots$) will denote positive constants.

Throughout the present paper, the sequence $\{m_n\}$ ($n = 1, 2, \dots$) defined by $m_1 = 1$ and

$$(0.9) \quad m_n = [e^{n/\log n}], \quad n = 2, 3, \dots,$$

will play a fundamental rôle. The fact that we adopt the sequence $\{m_n\}$ ($n = 1, 2, \dots$) instead of $\{a^n\}$ ($n = 1, 2, \dots$), which was used by A. Khintchine and A. Kolmogoroff, is essential in our proof, and will enable us to obtain our sharper results. The following inequalities, which are easy to prove, will be used very often:

$$(0.10) \quad m_n < m_{n+1} < c_3 m_n,$$

$$(0.11) \quad c_4 \frac{m_n}{\log \log m_n} < m_{n+1} - m_n < c_5 \frac{m_n}{\log \log m_n}$$

$$(0.12) \quad c_6 \left(\frac{m_n}{\log \log m_n} \right)^{\frac{1}{2}} < (m_{n+1} \log \log m_{n+1})^{\frac{1}{2}} \\ - (m_n \log \log m_n)^{\frac{1}{2}} < c_7 \left(\frac{m_n}{\log \log m_n} \right)^{\frac{1}{2}}.$$

It is not difficult to extend our results to the case in which the parameter n is continuous, i.e. the case of Brownian motion.⁴ We can define the upper and

⁴ Cf. A. Khintchine, loc. cit. 2. Cf. also N. Wiener, *Differential space*, Journal of Math. and Phys. 2 (1923), 131-174, and the book of P. Lévy quoted in footnote 5.

the lower classes in this case, and can obtain the corresponding results. It was stated by P. Lévy⁵ that A. Kolmogoroff has proved the following result: Let $\psi(\lambda) = \varphi(\lambda)/\lambda^{\frac{1}{2}}$ be monotone increasing. Then a necessary and sufficient condition that $\varphi(\lambda)$ belong to the lower class is given by the divergence of the integral

$$(0.13) \quad \int_0^\infty \psi(\lambda) e^{-2(\psi(\lambda))^2} \frac{d\lambda}{\lambda}.$$

It is easy to see that this is equivalent to Theorem 4. As far as I know, the proof of A. Kolmogoroff has not been published. Recently, J. Ville⁶ proved that the divergence of (0.13) is necessary. This corresponds to a special case of Theorem 1, but his proof is entirely different from ours.

1

THEOREM 1. $\varphi(n)$ belongs to the upper class if it is monotone increasing and if

$$(1.1) \quad \sum_{n=1}^{\infty} \Pr(A_{m_n}(\varphi(m_n))) < \infty.$$

PROOF. First we remark that we may assume that

$$(1.2) \quad \varphi(n) \leq (n \log \log n)^{\frac{1}{2}}$$

for sufficiently large n . Indeed, otherwise we may consider $\varphi_1(n) = \min(\varphi(n), (n \log \log n)^{\frac{1}{2}})$ instead of $\varphi(n)$. It is clear that $\varphi_1(n)$ is monotone increasing, that $\varphi_1(n)$ satisfies (1.1) if $\varphi(n)$ does (because, by (0.7), $\varphi_0(n) = (n \log \log n)^{\frac{1}{2}}$ satisfies (1.1)); and that if $\varphi_1(n)$ belongs to the upper class so does $\varphi(n)$ too.

Next we notice that, under the assumption (1.2), we have

$$(1.3) \quad \Pr(A_{m_{n+1}}(\varphi(m_n))) < c_8 \Pr(A_{m_n}(\varphi(m_n))).$$

This is an easy consequence of the relations (0.7), (0.10) and (0.11). We omit the proof.

Now assume that Theorem 1 is not true. Then there exists a constant $c_9 > 0$ such that, for any $M_0 = m_{n_0}$, there exists an $N_0 = m_{n'_0}$ ($n'_0 > n_0$) such that

$$(1.4) \quad \Pr\left(\sum_{M_0 < u \leq N_0} A_u(\varphi(u))\right) > c_9 > 0.$$

Let us put

$$(1.5) \quad \begin{aligned} B(u) &= A_u(\varphi(u)) - A_u(\varphi(u)) \sum_{M_0 < v < u} A_v(\varphi(v)) \\ &= E[t: f_u(t) > \varphi(u); f_v(t) \leq \varphi(v), M_0 < v < u]. \end{aligned}$$

⁵ P. Lévy, *Théorie de l'addition des variables aléatoires*, Paris, 1937.

⁶ J. Ville, *Étude critique de la notion de collectif*, Paris, 1937.

Then $\{B(u)\}$ ($M_0 < u \leq N_0$) are mutually disjoint, and

$$(1.6) \quad \sum_{M_0 < u \leq N_0} B(u) = \sum_{M_0 < u \leq N_0} A_u(\varphi(u)).$$

For each u ($M_0 < u \leq N_0$) take an n ($n_0 \leq n < n'_0$) such that $m_n < u \leq m_{n+1}$, and put

$$(1.7) \quad \Delta_{u, m_{n+1}}^+ = E[t: f_{m_{n+1}}(t) - f_u(t) \geq 0].$$

Then it is clear that $B(u)$ and $\Delta_{u, m_{n+1}}^+$ are independent, and hence

$$(1.8) \quad \Pr(B(u) \cdot \Delta_{u, m_{n+1}}^+) = \Pr(B(u))\Pr(\Delta_{u, m_{n+1}}^+) \geq \frac{1}{2} \cdot \Pr(B(u)).$$

On the other hand, since $t \in B(u) \cdot \Delta_{u, m_{n+1}}^+$ implies $f_{m_{n+1}}(t) \geq f_u(t) > \varphi(u) \geq \varphi(m_n)$, we have

$$(1.9) \quad B(u) \cdot \Delta_{u, m_{n+1}}^+ \subset A_{m_{n+1}}(\varphi(m_n))$$

for $m_n < u \leq m_{n+1}$. Hence, since $\{B(u) \cdot \Delta_{u, m_{n+1}}^+\}$ ($M_0 < u \leq N_0$) are mutually disjoint, we have, by (1.3), (1.8), (1.6) and (1.4),

$$(1.10) \quad \begin{aligned} c_8 \sum_{M_0 < u \leq N_0} \Pr(A_{m_n}(\varphi(m_n))) &\geq \sum_{M_0 < u \leq N_0} \Pr(A_{m_{n+1}}(\varphi(m_n))) \\ &\geq \sum_{M_0 < u \leq N_0} \Pr(B(u) \cdot \Delta_{u, m_{n+1}}^+) \geq \frac{1}{2} \sum_{M_0 < u \leq N_0} \Pr(B(u)) \\ &\geq \frac{1}{2} \Pr\left(\sum_{M_0 < u \leq N_0} B(u)\right) = \frac{1}{2} \Pr\left(\sum_{M_0 < u \leq N_0} A_u(\varphi(u))\right) > \frac{c_9}{2} > 0. \end{aligned}$$

Since c_8 and c_9 are positive constants, and since $M_0 = m_{n_0}$ can be arbitrarily large, this contradicts to the assumption (1.1). This proves Theorem 1.

COROLLARY 1. $\varphi(n) = (1/(2)^\frac{1}{2} + \epsilon)$ ($n \log \log n$) $^\frac{1}{2}$ belongs to the upper class for $\epsilon > 0$.

COROLLARY 2. The expression (0.6) belongs to the upper class for $\epsilon > 0$.

PROOF. Follows immediately from Theorem 1 and (0.7).

2

THEOREM 2. If $\varphi(n)$ is monotone increasing, then a necessary and sufficient condition that $\varphi(n)$ belong to the lower class is that, for almost all t , there exist infinitely many n such that

$$(2.1) \quad f_{m_n}(t) > \varphi(m_n).$$

PROOF. The sufficiency is obvious. In order to prove the necessity, let us assume that $\varphi(n)$ belongs to the lower class. First we remark that we may assume

$$(2.2) \quad \varphi(n) \leq (n \log \log n)^\frac{1}{2}$$

for sufficiently large n . Indeed, by Corollary 1 to Theorem 1, $\varphi_0(n) = (n \log \log n)^\frac{1}{2}$ belongs to the upper class. Hence, if we put $\varphi_1(n) = \min(\varphi(n), \varphi_0(n))$, then

$\varphi_1(n)$ belongs to the lower class if $\varphi(n)$ does; and if the necessity of the condition is proved for $\varphi_1(n)$, then it is obviously true for $\varphi(n)$ too.

By assumption, there exists a constant $c_{10} > 0$ such that for any $M_0 = m_{n_0}$ there exists an $N_0 = m_{n'_0}$ ($n'_0 > n_0$) such that

$$(2.3) \quad \Pr\left(\sum_{M_0 < u \leq N_0} A_u(\varphi(u))\right) > c_{10}.$$

Let us put

$$(2.4) \quad \begin{aligned} C(u) &= A_u(\varphi(u)) - A_u(\varphi(u)) \cdot \sum_{u < v \leq N_0} A_v(\varphi(v)) \\ &= E[t: f_u(t) > \varphi(u); f_v(t) \leq \varphi(v), u < v \leq N_0]. \end{aligned}$$

Then $\{C(u)\}$ ($M_0 < n \leq N_0$) are mutually disjoint, and

$$(2.5) \quad \sum_{M_0 < u \leq N_0} C(u) = \sum_{M_0 < u \leq N_0} A_u(\varphi(u)).$$

For each u ($M_0 < u \leq N_0$) take an n ($n_0 \leq n < n'_0$) such that $m_n < u \leq m_{n+}$ and put

$$(2.6) \quad \Delta_{m_n, u}^- = E[t: f_u(t) - f_{m_n}(t) \leq 0].$$

It is to be noticed that $C(u)$ and $\Delta_{m_n, u}^-$ are not independent, but it can be shown by computations⁷ that there exists a constant $c_{11} > 0$ such that

$$(2.7) \quad \Pr(C(u) \cdot \Delta_{m_n, u}^-) > c_{11} \Pr(C(u)).$$

⁷ We sketch the proof of (2.7): Let us put

$$C(u, k) = E[t: f_u(t) = k; f_v(t) \leq \varphi(v), u < v \leq N_0],$$

where $k > \varphi(u)$ is an integer or integer + $\frac{1}{2}$ according as u is even or odd. Then a simple calculation with binomial coefficients shows that

$$\Pr\left(\sum_{\varphi(u) < k \leq \varphi(u) + u/\varphi(u)} C(u, k)\right) > c_{48} \Pr(C(u)).$$

Thus it suffices to show that, for $\varphi(u) < k \leq \varphi(u) + u/\varphi(u)$,

$$\Pr(C(u, k) \cdot \Delta_{m_n, u}^-) > c_{47} \Pr(C(u, k)).$$

Now, it is easy to see that

$$\frac{\Pr(C(u, k))}{\Pr(C(u, k) \cdot \Delta_{m_n, u}^-)} < c_{48} \left(\frac{u}{\frac{u}{2} + k} \right) / \left(\frac{m_n}{\frac{u}{2} + k} \right);$$

and a simple calculation shows that

$$\left(\frac{u}{\frac{u}{2} + k} \right) > c_{49} \left(\frac{m_n}{\frac{u}{2} + k} \right),$$

which completes the proof of (2.7).

On the other hand, since $t \in C(u) \cdot \Delta_{m_n, u}^-$ implies $f_{m_n}(t) \geq f_u(t) > \varphi(u) \geq \varphi(m_n)$, we have

$$(2.8) \quad C(u) \cdot \Delta_{m_n, u}^- \subset A_{m_n}(\varphi(m_n))$$

for $m_n < u \leq m_{n+1}$. Hence, since $\{C(u) \cdot \Delta_{m_n, u}^-\} (M_0 < u \leq N_0)$ are mutually disjoint, we have, by (2.7) and (2.5),

$$(2.9) \quad \begin{aligned} \sum_{M_0 < m_n \leq N_0} Pr(A_{m_n}(\varphi(m_n))) &\geq \sum_{M_0 < u \leq N_0} Pr(C(u) \cdot \Delta_{m_n, u}^-) \\ &\geq c_{11} \sum_{M_0 < u \leq N_0} Pr(C(u)) = c_{11} Pr\left(\sum_{M_0 < u \leq N_0} C(u)\right) \\ &= c_{11} Pr\left(\sum_{M_0 < u \leq N_0} A_u(\varphi(u))\right) > c_{10} \cdot c_{11} > 0. \end{aligned}$$

Since c_{10} and c_{11} are absolute positive constants, and since $M_0 = m_{n_0}$ can be taken arbitrarily large, this means that the set of all t for which the inequality (2.1) holds for infinitely many n , has positive measure. By the law of 0 or 1, this set must have measure 1, and thus Theorem 2 is proved.

3

THEOREM 3. *Let $\varphi(n)$ be monotone increasing and let us assume that*

$$(3.1) \quad \varphi(m_{n+1}) - \varphi(m_n) > c_{12} (m_n / \log \log m_n)^{\frac{1}{2}}.$$

Then a necessary and sufficient condition that $\varphi(n)$ belong to the lower class is that

$$(3.2) \quad \sum_{n=1}^{\infty} Pr(A_{m_n}(\varphi(m_n))) = \infty.$$

PROOF. The necessity follows from Theorem 1, without assuming (3.1). In order to prove that the condition (3.1) is sufficient, let us assume that $\varphi(n)$ is monotone increasing and satisfies (3.1) and (3.2). We first notice that (3.1) and (0.12) imply

$$(3.3) \quad \varphi(m_{n+1}) - \varphi(m_n) > c_{13} ((m_{n+1} \log \log m_{n+1})^{\frac{1}{2}} - (m_n \log \log m_n)^{\frac{1}{2}}),$$

and hence

$$(3.4) \quad \varphi(m_n) > c_{14} (m_n \log \log m_n)^{\frac{1}{2}}.$$

From (3.4) and (0.7) it follows easily that

$$(3.5) \quad \lim_{n \rightarrow \infty} Pr(A_{m_n}(\varphi(m_n))) = 0.$$

Next we notice that we may assume

$$(3.6) \quad \varphi(n) \leq (n \log \log n)^{\frac{1}{2}}$$

for sufficiently large n . Indeed, otherwise we may consider $\varphi_1(n) = \min(\varphi(n), (n \log \log n)^{\frac{1}{2}})$ instead of $\varphi(n)$. Since $\varphi_0(n) = (n \log \log n)^{\frac{1}{2}}$ clearly satisfies (3.1), $\varphi_1(n)$ satisfies it too. Further, it is obvious that (3.2) is satisfied by $\varphi_1(n)$ whenever it is satisfied by $\varphi(n)$. Moreover, since $\varphi_0(n)$ belongs to the

upper class, by the corollary to Theorem 1, $\varphi_1(n)$ belongs to the lower class at the same time as $\varphi(n)$.

Because of the law of 0 or 1, and because of Theorem 2, it is sufficient to prove that there exists a constant $c_{15} > 0$ such that there exists, for any $M_0 = m_{n_0}$, an $N_0 = m_{n'_0}$ ($n'_0 > n_0$) such that

$$(3.7) \quad \Pr\left(\sum_{M_0 < m_n \leq N_0} A_{m_n}(\varphi(m_n))\right) > c_{15}.$$

Let $\delta > 0$ be a small positive number, which we shall determine later. Then, by (3.2) and (3.4), there exists an N such that, for any $M_0 = m_{n_0} > N$, an $N_0 = m_{n'_0}$ ($n'_0 > n_0$) exists such that

$$(3.8) \quad \delta < \sum_{M_0 < m_n \leq N_0} \Pr(A_{m_n}(\varphi(m_n))) < 2\delta.$$

We shall prove that if δ is chosen sufficiently small (but fixed), then (3.7) is satisfied, with the same integers M_0 and N_0 as in (3.8), by a suitable positive constant $c_{15} > 0$.

In order to prove this, let us first put

$$(3.9) \quad \begin{aligned} D(m_n) &= A_{m_n}(\varphi(m_n)) - A_{m_n}(\varphi(m_n)) \cdot \sum_{m_n < m_{n+r} \leq N_0} A_{m_{n+r}}(\varphi(m_{n+r})) \\ &= E[t: f_{m_n}(t) > \varphi(m_n); f_{m_{n+r}}(t) \leq \varphi(m_{n+r}), m_n < m_{n+r} \leq N_0]. \end{aligned}$$

Then $\{D(m_n)\}$ ($M_0 < m_n \leq N_0$) are mutually disjoint, and

$$(3.10) \quad \sum_{M_0 < m_n \leq N_0} D(m_n) = \sum_{M_0 < m_n \leq N_0} A_{m_n}(\varphi(m_n)).$$

Let us further put

$$(3.11) \quad \begin{aligned} D_1(m_n) &= A_{m_n}(\varphi(m_n)) - A_{m_n}\left(\varphi(m_n) + \frac{c_{12}}{2} \left(\frac{m_n}{\log \log m_n}\right)^{\frac{1}{2}}\right) \\ &= E\left[t: \varphi(m_n) < f_{m_n}(t) \leq \varphi(m_n) + \frac{c_{12}}{2} \left(\frac{m_n}{\log \log m_n}\right)^{\frac{1}{2}}\right]. \end{aligned}$$

Then a simple computation will show that⁸

⁸ We have clearly

$$\frac{\Pr(D_1(m_n))}{\Pr(A_{m_n}(\varphi(m_n)))} = \frac{\sum' \binom{m_n}{u}}{\sum_{u > \varphi(m_n)} \binom{m_n}{u}}$$

where the dash indicates that u runs only over the interval

$$\varphi(m_n), \quad \varphi(m_n) + \frac{c_{12}}{2} \left(\frac{m_n}{\log \log m_n}\right)^{\frac{1}{2}}.$$

A simple calculation shows that

$$\sum' \binom{m_n}{u} / \sum_{u > \varphi(m_n)} \binom{m_n}{u} > c_{16}$$

which proves (3.12).

$$(3.12) \quad \Pr(D_1(m_n)) > c_{16} \Pr(A_{m_n}(\varphi(m_n))).$$

Let us put

$$(3.13) \quad D_2(m_n) = D_1(m_n) \cdot E[t: f_{m_{n+r-1}}(t) - f_{m_{n+r}}(t) \leq 0, r = 1, 2, \dots, h],$$

where h is a positive integer which we shall determine later. Then it is easy to see that

$$(3.14) \quad \Pr(D_2(m_n)) \geq 2^{-h} \Pr(D_1(m_n)),$$

and that $t \in D_2(m_n)$ implies

$$(3.15) \quad f_{m_{n+r}}(t) \leq f_{m_n}(t) < \varphi(m_{n+r}),$$

for $r = 1, 2, \dots, h$. Let us further put

$$(3.16) \quad D_3(m_n) = D_2(m_n) \cdot E[t: f_{m_{n+r}}(t) \leq \varphi(m_{n+r}), m_{n+h} < m_{n+r} \leq N_0].$$

Then it is clear that $D_3(m_n) \subset D(m_n) \subset A_{m_n}(\varphi(m_n))$. In order to complete the proof of Theorem 3, it is sufficient to prove that, if δ is chosen sufficiently small and if h is chosen sufficiently large (but both fixed), then there exists a constant $c_{17} > 0$ such that

$$(3.17) \quad \Pr(D_3(m_n)) > c_{17} \Pr(A_{m_n}(\varphi(m_n))).$$

Indeed, (3.17) will imply

$$\begin{aligned} \Pr\left(\sum_{M_0 < m_n \leq N_0} A_{m_n}(\varphi(m_n))\right) &= \Pr\left(\sum_{M_0 < m_n \leq N_0} D(m_n)\right) \\ (3.18) \quad &= \sum_{M_0 < m_n \leq N_0} \Pr(D(m_n)) \geq \sum_{M_0 < m_n \leq N_0} \Pr(D_3(m_n)) \\ &> c_{17} \sum_{M_0 < m_n \leq N_0} \Pr(A_{m_n}(\varphi(m_n))) > c_{17} \cdot \delta, \end{aligned}$$

which means that (3.7) is satisfied by $c_{15} = c_{17} \cdot \delta > 0$, thus completing the proof of Theorem 3.

The rest of the proof of Theorem 3 is devoted to establishing the relation (3.17). For this purpose, put

$$(3.19) \quad D_{3,r}(m_n) = D_2(m_n) \cdot A_{m_{n+r}}(\varphi(m_{n+r})),$$

for all integers r such that $m_{n+h} < m_{n+r} \leq N_0$. It is easy to see that

$$(3.20) \quad D_2(m_n) \subset D_3(m_n) + \sum_{m_{n+h} < m_{n+r} \leq N_0} D_{3,r}(m_n).$$

We shall evaluate $\Pr(\sum_{m_{n+h} < m_{n+r} \leq N_0} D_{3,r}(m_n))$ by decomposing the sum into three parts: $\sum_{m_{n+h} < m_{n+r} \leq 2m_n}$, $\sum_{2m_n < m_{n+r} \leq m_n \log m_n}$, and $\sum_{m_n \log m_n < m_{n+r} \leq N_0}$.

In the first place, $t \in D_{3,r}(m_n)$ implies

$$\begin{aligned}
 f_{m_{n+r}}(t) - f_{m_{n+h}}(t) &\geq f_{m_{n+r}}(t) - f_{m_n}(t) \\
 &> \varphi(m_{n+r}) - \varphi(m_n) - \frac{c_{12}}{2} \left(\frac{m_n}{\log \log m_n} \right)^{\frac{1}{2}} \\
 (3.21) \quad &> \sum_{k=0}^{r-1} c_{12} \left(\frac{m_{n+k}}{\log \log m_{n+k}} \right)^{\frac{1}{2}} - \frac{c_{12}}{2} \left(\frac{m_n}{\log \log m_n} \right)^{\frac{1}{2}} \\
 &> \frac{c_{12} r}{2} \left(\frac{m_n}{\log \log m_n} \right)^{\frac{1}{2}}.
 \end{aligned}$$

Hence

$$(3.22) \quad \Pr(D_{3,r}(m_n)) \leq \alpha_r \cdot \Pr(D_2(m_n)),$$

$$\text{where } \alpha_r = \Pr \left(E \left[t: f_{m_{n+r}}(t) - f_{m_{n+h}}(t) > \frac{c_{12} r}{2} \left(\frac{m_n}{\log \log m_n} \right)^{\frac{1}{2}} \right] \right)$$

$$\begin{aligned}
 (3.23) \quad &= \Pr \left(A_{m_{n+r}-m_{n+h}} \left(\frac{c_{12} r}{2} \left(\frac{m_n}{\log \log m_n} \right)^{\frac{1}{2}} \right) \right) \\
 &< c_2 \frac{(m_{n+r} - m_{n+h})^{\frac{1}{2}}}{\frac{c_{12} r}{2} \left(\frac{m_n}{\log \log m_n} \right)^{\frac{1}{2}}} \exp \left[- \frac{2 \left(\frac{c_{12} r}{2} \right)^2 \frac{m_n}{\log \log m_n}}{m_{n+r} - m_{n+h}} \right].
 \end{aligned}$$

Since, on the other hand, $m_{n+h} < m_{n+r} \leq 2m_n$ implies

$$\begin{aligned}
 (3.24) \quad m_{n+r} - m_{n+h} &\leq \sum_{k=0}^{r-1} (m_{n+k+1} - m_{n+k}) \\
 &< c_5 r \frac{m_{n+r}}{\log \log m_{n+r}} \leq 2c_5 r \frac{m_n}{\log \log m_n},
 \end{aligned}$$

we have, by (0.7),

$$(3.25) \quad \alpha_r < c_{18} e^{-c_{19} r}$$

for $m_{n+h} < m_{n+r} \leq 2m_n$. Consequently,

$$(3.26) \quad \Pr \left(\sum_{m_{n+h} < m_{n+r} \leq 2m_n} D_{3,r}(m_n) \right) < c_{18} \cdot \Pr(D_2(m_n)) \cdot \sum_{r=h+1}^{\infty} e^{-c_{19} r}.$$

Secondly, $t \in D_{3,r}(m_n)$ and $2m_n < m_{n+r} \leq m_n \log m_n$ imply

$$\begin{aligned}
 (3.27) \quad f_{m_{n+r}}(t) - f_{m_{n+h}}(t) &> \frac{c_{12}}{2} \sum_{k=0}^{r-1} \left(\frac{m_{n+k}}{\log \log m_{n+k}} \right)^{\frac{1}{2}} \\
 &> \frac{c_{12}}{2c_7} \sum_{k=0}^{r-1} ((m_{n+k+1} \log \log m_{n+k+1})^{\frac{1}{2}} - (m_{n+k} \log \log m_{n+k})^{\frac{1}{2}}) \\
 &> c_{20} (m_{n+r} \log \log m_{n+r})^{\frac{1}{2}}.
 \end{aligned}$$

Hence

$$(3.28) \quad Pr(D_{3,r}(m_n)) < \beta_r \cdot Pr(D_2(m_n))$$

for $2m_n < m_{n+r} \leq m_n \log m_n$, where

$$(3.29) \quad \begin{aligned} \beta_r &= Pr(E[t: f_{m_{n+r}}(t) - f_{m_{n+h}}(t) > c_{20}(m_{n+r} \log \log m_{n+r})^{\frac{1}{2}}]) \\ &= Pr(A_{m_{n+r}-m_{n+h}}(c_{20}(m_{n+r} \log \log m_{n+r})^{\frac{1}{2}})) \\ &< c_2 \frac{(m_{n+r} - m_{n+h})^{\frac{1}{2}}}{c_{20}(m_{n+r} \log \log m_{n+r})^{\frac{1}{2}}} \exp \left[-\frac{2c_{20}^2 m_{n+r} \log \log m_{n+r}}{m_{n+r} - m_{n+h}} \right] \\ &< c_{21} e^{-c_{22} \log \log m_{n+r}} < \frac{c_{21}}{(\log m_n)^{c_{22}}}. \end{aligned}$$

On the other hand, the number of m_{n+r} 's satisfying $2m_n < m_{n+r} \leq m_n \log m_n$ does not exceed $c_{23}(\log \log m_n)^2$.⁹ Hence we have

$$(3.30) \quad Pr\left(\sum_{2m_n < m_{n+r} \leq m_n \log m_n} D_{3,r}(m_n)\right) < Pr(D_2(m_n)) \cdot c_{24} \frac{(\log \log m_n)^2}{(\log m_n)^{c_{21}}}.$$

Lastly, $t \in D_{3,r}(m_n)$ and $m_n \log m_n < m_{n+r} \leq N_0$ imply

$$(3.31) \quad \begin{aligned} f_{m_{n+r}}(t) - f_{m_{n+h}}(t) &> \varphi(m_{n+r}) - \varphi(m_n) - \frac{c_{12}}{2} \left(\frac{m_n}{\log \log m_n} \right)^{\frac{1}{2}} \\ &> \varphi(m_{n+r}) - 2(m_n \log \log m_n)^{\frac{1}{2}}. \end{aligned}$$

Hence

$$(3.32) \quad Pr(D_{3,r}(m_n)) < \gamma_r \cdot Pr(D_2(m_n))$$

for $m_n \log m_n < m_{n+r} \leq N_0$, where

$$(3.33) \quad \begin{aligned} \gamma_r &= Pr(E[t: f_{m_{n+r}}(t) - f_{m_{n+h}}(t) > \varphi(m_{n+r}) - 2(m_n \log \log m_n)^{\frac{1}{2}}]) \\ &= Pr(A_{m_{n+r}-m_{n+h}}(\varphi(m_{n+r}) - 2(m_n \log \log m_n)^{\frac{1}{2}})) \\ &< c_2 \frac{(m_{n+r} - m_{n+h})^{\frac{1}{2}}}{\varphi(m_{n+r}) - 2(m_n \log \log m_n)^{\frac{1}{2}}} \\ &\quad \cdot \exp \left[-\frac{2(\varphi(m_{n+r}) - 2(m_n \log \log m_n)^{\frac{1}{2}})^2}{m_{n+r} - m_{n+h}} \right] \\ &< c_2 \frac{(m_{n+r})^{\frac{1}{2}}}{\varphi(m_{n+r}) - 2(m_n \log \log m_n)^{\frac{1}{2}}} \\ &\quad \cdot \exp \left[-\frac{2(\varphi(m_{n+r}))^2}{m_{n+r}} + \frac{8\varphi(m_{n+r})(m_n \log \log m_n)^{\frac{1}{2}}}{m_{n+r}} \right]. \end{aligned}$$

⁹ It follows from (0.11) that the number of m_n 's in the interval $(x, 2x)$ does not exceed $c_{30} \log \log x$. Thus the number of m_n 's in the interval $(x, x \log x)$ does not exceed

$$c_{30} \log \log x \frac{\log \log x}{\log 2} < c_{33}(\log \log x)^2.$$

On the other hand, for sufficiently large n , $m_{n+r} > m_n \log m_n$ implies

$$(3.34) \quad \varphi(m_{n+r}) - 2(m_n \log \log m_n)^{\frac{1}{2}} > \frac{1}{2}\varphi(m_{n+r}),$$

$$(3.35) \quad \varphi(m_{n+r})(m_n \log \log m_n)^{\frac{1}{2}} < m_{n+r}.$$

Hence

$$(3.36) \quad \begin{aligned} \gamma_r &< c_{25} \cdot \frac{(m_{n+r})^{\frac{1}{2}}}{\varphi(m_{n+r})} \exp \left[-\frac{2(\varphi(m_{n+r}))^2}{m_{n+r}} \right] \\ &< c_{26} \Pr(A_{m_{n+r}}(\varphi(m_{n+r}))). \end{aligned}$$

Consequently,

$$(3.37) \quad \begin{aligned} \Pr \left(\sum_{m_n \log m_n < m_{n+r} \leq N_0} D_{3,r}(m_n) \right) \\ &< \Pr(D_2(m_n)) \cdot c_{25} \sum_{m_0 < m_{n+r} \leq N_0} \Pr(A_{m_{n+r}}(\varphi(m_{n+r}))) \\ &< c_{26} \cdot 2\delta \cdot \Pr(D_2(m_n)). \end{aligned}$$

Combining (3.26), (3.30) and (3.37), we have finally

$$(3.38) \quad \begin{aligned} \Pr \left(\sum_{m_n+h < m_{n+r} \leq N_0} D_{3,r}(m_n) \right) \\ &< \Pr(D_2(m_n)) \left\{ c_{18} \cdot \sum_{r=h+1}^{\infty} e^{-c_{19}r} + c_{24} \frac{(\log \log m_n)^2}{(\log m_n)^{c_{22}}} + c_{26} \cdot 2\delta \right\}. \end{aligned}$$

Hence, if we take h sufficiently large and δ sufficiently small, then we have

$$(3.39) \quad \Pr \left(\sum_{m_n+h < m_{n+r} \leq N_0} D_{3,r}(m_n) \right) < \theta \cdot \Pr(D_2(m_n)),$$

where θ is a constant with $0 < \theta < 1$. Consequently, by (3.20),

$$(3.40) \quad \begin{aligned} \Pr(D_3(m_n)) &> (1 - \theta) \cdot \Pr(D_2(m_n)) \\ &> 2^{-h}(1 - \theta) \cdot \Pr(D_1(m_n)) > (1 - \theta) \cdot c_{27} \Pr(A_{m_n}(\varphi(m_n))) \text{ (by 3.12)} \\ &> c_{17} \Pr(A_{m_n}(\varphi(m_n))), \end{aligned}$$

which proves (3.17). The proof of Theorem 3 is completed.

COROLLARY 1. $\varphi(n) = (1/\sqrt{2} + \epsilon)(n \log \log n)^{\frac{1}{2}}$ belongs to the lower class for $\epsilon \leq 0$.

COROLLARY 2. The expression (0.6) belongs to the lower class for $\epsilon \leq 0$.

PROOF. Follows immediately from Theorem 3 and (0.7).

4

THEOREM 4. Let $\varphi(n)/n^{\frac{1}{2}}$ be monotone increasing. Then a necessary and sufficient condition that $\varphi(n)$ belong to the lower class is that

$$(4.1) \quad \sum_{n=1}^{\infty} \Pr(A_{m_n}(\varphi(m_n))) = \infty.$$

We need the following

LEMMA 1. Let $M_1 < N_1 < M_2 < N_2 < \dots < M_i < N_i < \dots$ be a sequence of positive integers tending to infinity, and let $\varphi(n)$ be such that

$$(4.2) \quad \varphi(m_{n+1}) - \varphi(m_n) > c_{28} \left(\frac{m_n}{\log \log m_n} \right)^{\frac{1}{2}} \quad M_i < m_n \quad m_{n+1} \leq N_i,$$

$$(4.3) \quad \varphi(m_n) > c_{29} (m_n \log \log m_n)^{\frac{1}{2}}, \quad M_i < m_n < m_{n+1} \leq N_i,$$

$$(4.4) \quad \sum_{M_i < m_n \leq N_i} \Pr(A_{m_n}(\varphi(m_n))) > c_{30}.$$

Then $\varphi(n)$ belongs to the lower class.

We do not give the proof of Lemma 1, since it can be carried out in the same way as in Theorem 3.

PROOF OF THEOREM 4. The necessity of the condition is clear by Theorem 1. In order to prove that it is sufficient, let us assume that $\varphi(n)/n^{\frac{1}{2}}$ is monotone increasing and that (4.1) is satisfied. We shall prove that there exists a sequence of integers $M_1 < N_1 < M_2 < N_2 < \dots < M_i < N_i < \dots$ tending to infinity, which satisfies the conditions of Lemma 1.

If we have

$$(4.5) \quad \varphi(m_n) < \frac{1}{16} (m_n \log \log m_n)^{\frac{1}{2}}$$

for all sufficiently large n , then the fact that $\varphi_0(n) = \frac{1}{16} (n \log \log n)^{\frac{1}{2}}$ belongs to the lower class (see Corollary 1 to Theorem 3), together with Theorem 2, will imply that $\varphi(n)$ belongs to the lower class. On the other hand, if

$$(4.6) \quad \varphi(m_n) > \frac{1}{32} (m_n \log \log m_n)^{\frac{1}{2}}$$

for sufficiently large n , then

$$(4.7) \quad \varphi(m_{n+1}) \geq \left(\frac{m_{n+1}}{m_n} \right)^{\frac{1}{2}} \varphi(m_n) > \left(1 + \frac{c_{31}}{\log \log m_n} \right) \varphi(m_n)$$

by (0.11), and hence

$$(4.8) \quad \varphi(m_{n+1}) - \varphi(m_n) > c_{31} \frac{\varphi(m_n)}{\log \log m_n} > \frac{c_{31}}{20} \left(\frac{m_n}{\log \log m_n} \right)^{\frac{1}{2}}.$$

Consequently, by Theorem 3, $\varphi(n)$ must belong to the lower class again.

Thus, in order to prove Theorem 4, we have only to consider the case when there exist two sequences of integers tending to infinity $\{M_i\} = \{m_{n_i}\}$ ($i = 1, 2, \dots$) and $\{N_i\} = \{m_{n'_i}\}$ ($i = 1, 2, \dots$) such that $M_1 < N_1 < M_2 < N_2 < \dots < M_i < N_i < \dots$, and

$$(4.9) \quad \varphi(M_i) = \varphi(m_{n_i}) \geq \frac{1}{16} (M_i \log \log M_i)^{\frac{1}{2}},$$

$$(4.10) \quad \varphi(N_i) = \varphi(m_{n'_i}) \leq \frac{1}{32} (N_i \log \log N_i)^{\frac{1}{2}}.$$

We may assume that

$$(4.11) \quad \varphi(m_n) < \frac{1}{16} (m_n \log \log m_n)^{\frac{1}{2}}$$

for $M_i < m_n \leq N_i$ (i.e. for $n_i < n \leq n'_i$) ($i = 1, 2, \dots$).

We shall prove that the conditions of Lemma 1 are all satisfied by these $\{M_i\}$ ($i = 1, 2, \dots$) and $\{N_i\}$ ($i = 1, 2, \dots$). Since $\varphi(M_i)/(M_i)^{\frac{1}{2}} \leq \varphi(N_i)/N_i^{\frac{1}{2}}$ by assumption, we have $\frac{1}{10}(\log \log M_i)^{\frac{1}{2}} \leq \frac{1}{10}(\log \log N_i)^{\frac{1}{2}}$ for $i = 1, 2, \dots$. Since $M_i, N_i \rightarrow \infty$ as $i \rightarrow \infty$, it follows that we have $M_i^{\frac{1}{2}} \leq N_i$ for sufficiently large i .

Let now $M_i < m_n < N_i$. Then

$$\begin{aligned}
 (4.12) \quad Pr(A_{m_n}(\varphi(m_n))) &> c_1 \frac{(m_n)^{\frac{1}{2}}}{\varphi(m_n)} e^{-2(\varphi(m_n))^{\frac{1}{2}}/m_n} \\
 &> c_1 \frac{10}{(\log \log m_n)^{\frac{1}{2}}} e^{-(\log \log m_n)/50} = \frac{10 c_1}{(\log \log m_n)^{\frac{1}{2}} (\log m_n)^{1/50}} \\
 &> \frac{1}{(\log m_n)^{1/49}}
 \end{aligned}$$

for sufficiently large i . Since $2 \cdot \log M_i \cdot \log \log M_i < n < 3 \cdot \log M_i \cdot \log \log M_i$ implies $\log M_i < n/\log n < 4 \log M_i$, or equivalently $M_i < e^{n/\log n} < M_i^4$, for sufficiently large i , we have

$$\begin{aligned}
 (4.13) \quad \sum_{M_i < m_n \leq N_i} Pr(A_{m_n}(\varphi(m_n))) &> \sum_{M_i < m_n \leq N_i} \frac{1}{(\log m_n)^{1/49}} \\
 &> \sum_{M_i < m_n \leq M_i^4} \frac{1}{(\log m_n)^{1/49}} > \sum_{2 p_i < n \leq 3 p_i} \frac{1}{n^{1/49}},
 \end{aligned}$$

where $p_i = \log N_i \cdot \log \log N_i$. Thus (4.4) is satisfied. (4.3) is clearly satisfied with $c_{29} = \frac{1}{20}$; (4.8) shows that (4.2) is also satisfied. This completes the proof of Theorem 4.

5

THEOREM 5. *Let $\varphi(n)$ satisfy*

$$(5.1) \quad \varphi(n) > c_{32}(n \log \log n)^{\frac{1}{2}},$$

$$(5.2) \quad \sum_{n=1}^{\infty} Pr(A_{m_n}(\varphi(m_n))) = \infty.$$

Then $\varphi(n)$ belongs to the lower class.

To prove Theorem 5 we need the following

LEMMA 2. *Let $\varphi(n)$ be monotone increasing, and let $\{m_{n_i}\}$ ($i = 1, 2, \dots$) be a subsequence of $\{m_n\}$ ($n = 1, 2, \dots$) such that*

$$(5.3) \quad \varphi(m_{n_{i+1}}) \geq \varphi(m_{n_i}) + c_{33}((m_{n_{i+1}} \log \log m_{n_{i+1}})^{\frac{1}{2}} - (m_{n_i} \log \log m_{n_i})^{\frac{1}{2}})$$

$$(5.4) \quad \sum_{i=1}^{\infty} Pr(A_{m_{n_i}}(\varphi(m_{n_i}))) = \infty.$$

Then $\varphi(n)$ belongs to the lower class.

Since the proof of Lemma 2 can be carried out exactly as in the proof of Theorem 3, we omit the proof.

PROOF OF THEOREM 5. As in the proof of Theorem 3, we may assume that

$$(5.5) \quad \varphi(n) \leq (n \log \log n)^{\frac{1}{2}}$$

for sufficiently large n . We shall find a subsequence $\{m_{n_i}\}$ ($i = 1, 2, \dots$) of $\{m_n\}$ ($n = 1, 2, \dots$) which satisfies the conditions of Lemma 2. For this purpose we classify the integers m_n into two classes. The first class I consists of all integers m_p for which

$$(5.6) \quad \varphi(m_q) \geq \varphi(m_p) + \epsilon((m_q \log \log m_q)^{\frac{1}{2}} - (m_p \log \log m_p)^{\frac{1}{2}})$$

for all $q \geq p$, where ϵ is a positive constant with $0 < \epsilon < c_{32}$ which we shall determine later. All other integers m_p will belong to the second class II. We shall prove that, if we denote by $\{m_{n_i}\}$ ($i = 1, 2, \dots$, $m_{n_i} < m_{n_{i+1}}$) the integers of the class I, then this sequence satisfies the conditions of Lemma 2. Indeed, (5.3) is clear from (5.6). In order to prove (5.4) for the m_{n_i} 's of the class I, let us denote by II_i the set of all integers m_p of the class II such that $m_p < m_{n_i}$ and

$$(5.7) \quad \varphi(m_{n_i}) < \varphi(m_p) + \epsilon((m_{n_i} \log \log m_{n_i})^{\frac{1}{2}} - (m_p \log \log m_p)^{\frac{1}{2}}).$$

By definition, for each m_p of the class II, there exists an m_q ($m_q > m_p$) such that

$$(5.8) \quad \varphi(m_q) < \varphi(m_p) + \epsilon((m_q \log \log m_q)^{\frac{1}{2}} - (m_p \log \log m_p)^{\frac{1}{2}}).$$

Because of (5.1) and the relation $\epsilon < c_{32}$, there exists, for each m_p of II, a largest integer m_q ($m_q > m_p$) satisfying (5.8). This m_q clearly belongs to I. Hence we have $\sum_{i=1}^{\infty} II_i = II$ (II_i are not necessarily mutually disjoint).

Thus in order to prove (5.4), we need only prove that there exists a constant $c_{34} > 0$ such that

$$(5.9) \quad \sum_{m_p \in II_i} Pr(A_{m_p}(\varphi(m_p))) < c_{34} Pr(A_{m_{n_i}}(\varphi(m_{n_i}))).$$

For this purpose we shall first show that

$$(5.10) \quad m_{n_i} < c_{35} m_p$$

for all $m_p \in II_i$, where c_{35} is independent of i and p . Indeed, if (5.10) is false, we have

$$\begin{aligned} \varphi(m_{n_i}) &< \varphi(m_p) + \epsilon((m_{n_i} \log \log m_{n_i})^{\frac{1}{2}} - (m_p \log \log m_p)^{\frac{1}{2}}) \\ (5.11) \quad &< (m_p \log \log m_p)^{\frac{1}{2}} + \epsilon(m_{n_i} \log \log m_{n_i})^{\frac{1}{2}} \\ &< \left(\frac{1}{\sqrt{c_{35}}} + \epsilon \right) (m_{n_i} \log \log m_{n_i})^{\frac{1}{2}}, \end{aligned}$$

and this is a contradiction to (5.1) if c_{35} is sufficiently large.

By (5.5), (5.7) and (5.10), if $m_p = m_{n_i-k} \in \Pi_i$, then we have

$$\begin{aligned}
 Pr(A_{m_p}(\varphi(m_p))) &< c_2 \frac{m_p^{\frac{1}{2}}}{\varphi(m_p)} e^{-2(\varphi(m_p))^2/m_p} \\
 &< \frac{c_2}{(\log \log m_p)^{\frac{1}{2}}} \\
 &\quad \cdot \exp \left[-\frac{2\{\varphi(m_{n_i}) - \epsilon((m_{n_i} \log \log m_{n_i})^{\frac{1}{2}} - (m_p \log \log m_p)^{\frac{1}{2}})\}^2}{m_{n_i}} \right] \\
 (5.12) \quad &< \frac{c_2}{(\log \log m_p)^{\frac{1}{2}}} \\
 &\quad \cdot \exp \left[-\frac{2(\varphi(m_{n_i}))^2 - 4\epsilon\varphi(m_{n_i})((m_{n_i} \log \log m_{n_i})^{\frac{1}{2}} - (m_p \log \log m_p)^{\frac{1}{2}})}{m_{n_i} - k \cdot c_5 \frac{m_p}{\log \log m_p}} \right] \\
 &< \frac{c_{36}}{(\log \log m_n)^{\frac{1}{2}}} \exp \left[-\frac{2(\varphi(m_{n_i}))^2}{m_{n_i}} \right] \cdot \eta^2 < c_{37} \cdot Pr(A_{m_{n_i}}(\varphi(m_{n_i}))) \cdot \eta^2
 \end{aligned}$$

where

$$\begin{aligned}
 \eta &= \exp \left[\frac{(\varphi(m_{n_i}))^2}{m_{n_i}} \right. \\
 &\quad \left. - \frac{(\varphi(m_{n_i}))^2 - 2\epsilon\varphi(m_{n_i})((m_{n_i} \log \log m_{n_i})^{\frac{1}{2}} - (m_p \log \log m_p)^{\frac{1}{2}})}{m_{n_i} - kc_{38} \frac{m_{n_i}}{\log \log m_{n_i}}} \right] \\
 &< \exp \left[\frac{(\varphi(m_{n_i}))^2}{m_{n_i}} \right. \\
 &\quad \left. - \frac{(\varphi(m_{n_i}))^2 - 2\epsilon\varphi(m_{n_i})((m_{n_i} \log \log m_{n_i})^{\frac{1}{2}} - (m_p \log \log m_p)^{\frac{1}{2}})}{m_{n_i}} \right] \\
 (5.13) \quad &\quad \cdot \left(1 + k \frac{c_{39}}{\log \log m_{n_i}} \right) \\
 &< \exp \left[\frac{2\epsilon\varphi(m_{n_i})}{m_{n_i}} ((m_{n_i} \log \log m_{n_i})^{\frac{1}{2}} - (m_p \log \log m_p)^{\frac{1}{2}}) \right. \\
 &\quad \left. - \frac{kc_{40}(\varphi(m_{n_i}))^2}{m_{n_i} \log \log m_{n_i}} \right] \\
 &< \exp \left[\frac{2\epsilon\varphi(m_{n_i})}{m_{n_i}} \cdot k \cdot c_7 \left(\frac{m_{n_i}}{\log \log m_{n_i}} \right)^{\frac{1}{2}} - k \frac{c_{40}(\varphi(m_{n_i}))^2}{m_{n_i} \log \log m_{n_i}} \right] \\
 &< \exp [(2\epsilon \cdot c_7 c_{32} - c_{40} \cdot c_{32}^2) \cdot k].
 \end{aligned}$$

Hence, if we take ϵ sufficiently small, then

$$(5.14) \quad \eta < e^{-c_{42}k}$$

with a positive constant c_{42} . Hence

$$(5.15) \quad \sum_{m_p \in \Pi_i} Pr(A_{m_p}(\varphi(m_p))) < c_{37} Pr(A_{m_{n_i}}(\varphi(m_{n_i}))) \cdot \sum_{k=1}^{\infty} e^{-2c_{42}k} \\ = \frac{c_{37}}{1 - e^{-2c_{42}}} Pr(A_{m_{n_i}}(\varphi(m_{n_i})))$$

which proves (5.9). This completes the proof of Theorem 5.

Before concluding this chapter, let us add some more results without proof.

1). If $\varphi(n)$ is monotone increasing and belongs to the lower class, then $\varphi(n) + c(n/\log \log n)^{\frac{1}{2}}$ belongs to the lower class for all c .

This result is the best possible. For, if $\psi(n) \rightarrow \infty$, then we can find a monotone function $\varphi(n)$ belonging to the lower class such that $\varphi(n) + \psi(n)(n/\log \log n)^{\frac{1}{2}}$ belongs to the upper class.

2). If $\varphi(n)$ is monotone increasing and belongs to the lower class, then $\varphi(n) + c(n/\varphi(n))$ belongs to the lower class for all c . Since we can always assume that $\varphi(n) < (n \log \log n)^{\frac{1}{2}}$, 2) is slightly stronger than 1).

3). Let $\varphi(n)$ be monotone increasing, and suppose that it belongs to the upper class. Then for almost all t , there exist only finitely many n such that for some $m < n$, $|f_n(t) - f_m(t)| > \varphi(n)$.

4). For almost all t , we have

$$(5.16) \quad \limsup_{n \rightarrow \infty} \frac{\sum_{k=1}^n f_k(t)}{\frac{1}{2} n^{\frac{1}{2}} \left(\frac{\log \log n}{2} \right)^{\frac{1}{2}}} = 1.$$

Professor J. L. Doob suggested that if $n_1 < n_2 < \dots$ is a sequence of integers with $n_{i+1}/n_i > c > 1$, then for almost all t ,

$$(5.17) \quad \lim_{p \rightarrow \infty} \frac{1}{p} \sum_{i=1}^p \frac{\sum_{k=1}^{n_i} f_k(t)}{n_i^{\frac{1}{2}}} = 0.$$

Indeed, it is not difficult to show that (5.17) holds. In fact, the condition $n_{i+1}/n_i > c > 1$ can be weakened, but it is necessary that n_i tends to infinity with a certain speed (quicker than i).

5). There exists a continuous strictly decreasing function $\psi(x)$ defined for $0 \leq x \leq 1$, with $\psi(0) = 1$, $\psi(1) = 0$, such that, for almost all t , the upper density of the set of n 's for which

$$(5.18) \quad f_n(t) \geq x \left(\frac{n \log \log n}{2} \right)^{\frac{1}{2}}$$

is exactly $\psi(x)$.¹⁰

¹⁰ This problem was suggested by W. Ambrose.

6

In this final chapter, we shall construct an increasing function $\varphi(n)$ such that

$$(6.1) \quad \sum_{n=1}^{\infty} \Pr(A_{m_n}(\varphi(m_n))) = \infty,$$

and nevertheless $\varphi(n)$ belongs to the upper class. This shows that the *converse of Theorem 1 is not true*.

We put

$$(6.2) \quad p_k = 2^{2^k}, \quad k = 1, 2, \dots,$$

and define $\varphi(n)$ as follows:

$$(6.3) \quad \varphi(n) = \log k \cdot \sqrt{p_k}, \quad \text{for } p_{k-1} < n \leq p_k.$$

It follows from (0.11) that the number of m_n 's satisfying $\frac{1}{2}p_k < m_n \leq p_k$ is $\geq c_{43} \log \log p_k$ and hence $\geq c_{44} 2^k$. Consequently, from (0.7) we have

$$(6.4) \quad \sum_{\frac{1}{2}p_k < m_n \leq p_k} \Pr(A_{m_n}(\varphi(m_n))) > \frac{c_1}{\log k} e^{-4(\log k)^2} \cdot c_{44} 2^k \geq c_{45} > 0.$$

Since this is true for each k , (6.1) is proved.

Denote now

$$(6.5) \quad M_k = E[t: \max_{1 \leq n \leq p_k} f_n(t) > \log k \cdot \sqrt{p_k}].$$

In order to show that $\varphi(n)$ belongs to the upper class, it is clearly sufficient to prove that

$$(6.6) \quad \sum_{k=1}^{\infty} \Pr(M_k) < \infty.$$

It is easy to see that¹¹

$$(6.7) \quad \Pr(M_k) \leq 2 \Pr(E[t: f_{p_k}(t) > \log k \sqrt{p_k}]).$$

¹¹ In general, we have

$$\Pr(E[t: \max_{1 \leq n \leq p} f_n(t) > x]) \leq 2 \Pr(E[t: f_p(t) > x]).$$

Indeed, we have

$$\begin{aligned} \Pr(E[t: \max_{1 \leq n \leq p} f_n(t) > x]) &= \Pr(E[t: f_p(t) > x]) \\ &+ \sum_{n=1}^{p-1} \Pr(E[t: f_1(t) \leq x, \dots, f_{n-1}(t) \leq x, f_n(t) > x, f_p(t) \leq x]) \\ &= \Pr(E[t: f_p(t) > x]) \\ &+ \sum_{n=1}^{p-1} \Pr(E[t: f_1(t) \leq x, \dots, f_{n-1}(t) \leq x, f_n(t) > x, f_p(t) \geq 2f_n(t) - x]) \\ &\leq \Pr(E[t: f_p(t) > x]) \\ &+ \sum_{n=1}^{p-1} \Pr(E[t: f_1(t) \leq x, \dots, f_{n-1}(t) \leq x, f_n(t) > x, f_p(t) > x]) \\ &= 2 \Pr(E[t: f_p(t) > x]). \end{aligned}$$

Thus from (0.7) we have

$$(6.8) \quad \sum_{k=1}^{\infty} Pr(M_k) \leq 2 \cdot c_2 \cdot \sum_{k=1}^{\infty} \frac{1}{\log k} e^{-2(\log k)^2} < \infty,$$

which proves (6.6).

My indebtedness to my friend S. Kakutani is very great. In fact, he wrote the whole paper after listening to my rough oral exposition.

UNIVERSITY OF PENNSYLVANIA
PHILADELPHIA, PA.

ON AN ELEMENTARY PROOF OF SOME ASYMPTOTIC FORMULAS IN THE THEORY OF PARTITIONS

By P. ERDÖS

(Received January 7, 1942)

Denote by $p(n)$ the number of partitions of n . Hardy and Ramanujan¹ proved in their classical paper that

$$p(n) \sim \frac{1}{4n^{3/4}} e^{cn^{1/4}}, \quad c = \pi(\frac{2}{3})^{1/2},$$

using complex function theory. The main purpose of the present paper is to give an elementary proof of this formula. But we can only prove with our elementary method that

$$(1) \quad p(n) \sim \frac{a}{n} e^{cn^{1/4}}$$

and are unable to prove that $a = 1/4.3^{1/4}$.

Our method will be very similar to that used in a previous paper.² The starting point will be the following identity:

$$(2) \quad np(n) = \sum_{v=1}^n \sum_{k=1}^{\infty} vp(n - kv), \quad p(0) = p(-m) = 0.$$

(We easily obtain (2) by adding up all the $p(n)$ partitions of n , and noting that v occurs in $p(n - v)$ partitions.) (2) is of course well known. In fact, Hardy and Ramanujan state in their paper³ that by using (2) they have obtained an elementary proof of

$$(3) \quad \log p(n) \sim cn^{1/4}.$$

The proof of (3) is indeed easy. First we show that

$$(4) \quad p(n) < e^{cn^{1/4}}.$$

We use induction. (4) clearly holds for $n = 1$. By (2) and the induction hypothesis we have

$$np(n) < \sum_{v=1}^n \sum_{k=1}^{\infty} ve^{c(n-kv)^{1/4}} < \sum_{v=1}^{\infty} \sum_{k=1}^{\infty} ve^{cn^{1/4} - ckv/2n^{1/4}} = e^{cn^{1/4}} \sum_{k=1}^{\infty} \frac{e^{-kc/2n^{1/4}}}{(1 - e^{kc/2n^{1/4}})^2}.$$

¹ Hardy, Ramanujan, *Asymptotic formulae in combinatory analysis*, Proc. London Math. Soc. 17, (1918), pp. 75-115.

² Erdős, *On some asymptotic formulas in the theory of factorisation numerorum*, these Annals 42, (1941), pp. 989-993.

³ Hardy, Ramanujan, *ibid*, p. 79.

Now it is easy to see that for all real x , $\frac{e^{-x}}{(1 - e^{-x})^2} < \frac{1}{x^2}$. Thus

$$np(n) < e^{cn^{\frac{1}{2}}} \sum_{k=1}^{\infty} \frac{4n}{c^2 k^2} = ne^{cn^{\frac{1}{2}}},$$

which proves (4).

Similarly but with slightly longer calculations, we can prove that for every $\epsilon > 0$ there exists an $A > 0$ such that

$$(5) \quad p(n) > \frac{1}{A} e^{(c-\epsilon)n^{\frac{1}{2}}}$$

(4) and (5) clearly imply (3).

To prove (1) we need the following

LEMMA 1:

$$(6) \quad \sum = \sum_{v=1}^{\infty} \sum_{\substack{k=1 \\ kv < n}}^{\infty} \frac{ve^{c(n-kv)^{\frac{1}{2}}}}{n - kv} = e^{cn^{\frac{1}{2}}} \left[1 + O\left(\frac{1}{n^{\frac{1}{2}+\epsilon}}\right) \right],$$

for some fixed $\epsilon > 0$.

PROOF. We omit as many details as possible, since the proof is quite straight forward and uninteresting. We evidently have by expanding $1/(n - kv)$ and omitting the terms with $kv > n^{\frac{1}{2}+\epsilon}$

$$\begin{aligned} \sum_{v=1}^n \sum_{\substack{k=1 \\ kv < n}}^n \frac{ve^{c(n-kv)^{\frac{1}{2}}}}{n - kv} &= \frac{1}{n} \sum_{v=1}^{\infty} \sum_{\substack{k=1 \\ kv < n}}^{\infty} ve^{c(n-kv)^{\frac{1}{2}}} + \frac{1}{n^2} \sum_{v=1}^{\infty} \sum_{\substack{k=1 \\ kv < n}}^{\infty} kv^2 e^{c(n-kv)^{\frac{1}{2}}} \\ &\quad + O\left(\frac{e^{cn^{\frac{1}{2}}}}{n^{\frac{1}{2}+\epsilon}}\right) = \sum_1 + \sum_2 + O\left(\frac{e^{cn^{\frac{1}{2}}}}{n^{\frac{1}{2}+\epsilon}}\right). \end{aligned}$$

Now

$$\sum_2 = \frac{e^{cn^{\frac{1}{2}}}}{n^2} \sum_{v=1}^{\infty} \sum_{k=1}^{\infty} kv^2 e^{-ckv/2n^{\frac{1}{2}}} + O\left(\frac{e^{cn^{\frac{1}{2}}}}{n^{\frac{1}{2}+\epsilon}}\right).$$

(It is easy to see that the other terms of $e^{c(n-kv)^{\frac{1}{2}}}$ can be neglected and that the summation for v and k can be extended to ∞ .) Thus

$$\begin{aligned} \sum_2 &= \frac{e^{cn^{\frac{1}{2}}}}{n^2} \sum_{k=1}^{\infty} \frac{2k}{(1 - e^{-kc/2n^{\frac{1}{2}}})^3} + O\left(\frac{e^{cn^{\frac{1}{2}}}}{n^{\frac{1}{2}+\epsilon}}\right) = e^{cn^{\frac{1}{2}}} \sum_{k=1}^{\infty} \frac{2k \cdot 8n^{\frac{1}{2}}}{k^3 c^3} \\ &\quad + O\left(\frac{e^{cn^{\frac{1}{2}}}}{n^{\frac{1}{2}+\epsilon}}\right) = \frac{4}{c} \frac{e^{cn^{\frac{1}{2}}}}{n^{\frac{1}{2}}} + O\left(\frac{e^{cn^{\frac{1}{2}}}}{n^{\frac{1}{2}+\epsilon}}\right). \end{aligned}$$

On the other hand

$$\begin{aligned}\Sigma_1 &= \frac{e^{cn^{\frac{1}{2}}}}{n} \sum_{v=1}^{\infty} \sum_{k=1}^{\infty} v e^{-ckv/2n^{\frac{1}{2}} - ck^2 v^2/8n^{\frac{1}{2}}} + O\left(\frac{e^{cn^{\frac{1}{2}}}}{n^{\frac{1}{2}+\epsilon}}\right) \\ &= \frac{e^{cn^{\frac{1}{2}}}}{n} \left(\sum_{v=1}^{\infty} \sum_{k=1}^{\infty} v e^{-ckv/2n^{\frac{1}{2}}} - \frac{ck^2 v^3}{8n^{\frac{1}{2}}} e^{-ckv/2n^{\frac{1}{2}}} \right) = \Sigma'_1 - \Sigma''_1 + O\left(\frac{e^{cn^{\frac{1}{2}}}}{n^{\frac{1}{2}+\epsilon}}\right). \\ \Sigma''_1 &= \frac{ce^{cn^{\frac{1}{2}}}}{8n^{\frac{1}{2}}} \sum_{v=1}^{\infty} \sum_{k=1}^{\infty} k^2 v^3 e^{-ckv/2n^{\frac{1}{2}}} = \frac{ce^{cn^{\frac{1}{2}}}}{8n^{\frac{1}{2}}} \sum_{k=1}^{\infty} \frac{6k^2}{(1 - e^{-ck/2n^{\frac{1}{2}}})^4} + O\left(\frac{e^{cn^{\frac{1}{2}}}}{n^{\frac{1}{2}+\epsilon}}\right) \\ &= \frac{ce^{cn^{\frac{1}{2}}}}{8n^{\frac{1}{2}}} \sum_{k=1}^{\infty} \frac{6k^2 \cdot 16n^2}{k^4 c^4} + O\left(\frac{e^{cn^{\frac{1}{2}}}}{n^{\frac{1}{2}+\epsilon}}\right) = \frac{3}{c} \frac{e^{cn^{\frac{1}{2}}}}{n^{\frac{1}{2}}} + O\left(\frac{e^{cn^{\frac{1}{2}}}}{n^{\frac{1}{2}+\epsilon}}\right). \\ \Sigma'_1 &= \frac{e^{cn^{\frac{1}{2}}}}{n} \sum_{v=1}^{\infty} \sum_{k=1}^{\infty} v e^{-ckv/2n^{\frac{1}{2}}} = \frac{e^{cn^{\frac{1}{2}}}}{n} \sum_{k=1}^{\infty} \frac{e^{-ck/2n^{\frac{1}{2}}}}{(1 - e^{-ck/2n^{\frac{1}{2}}})^2}.\end{aligned}$$

A simple calculation shows that

$$\frac{e^{-x}}{(1 - e^{-x})^2} = \frac{1}{x^2} + O(1), \quad \text{i.e.} \quad \frac{e^{-ck/2n^{\frac{1}{2}}}}{(1 - e^{-ck/2n^{\frac{1}{2}}})^2} = \frac{4n}{c^2 k^2} + O(1).$$

Hence

$$\Sigma'_1 = \frac{e^{cn^{\frac{1}{2}}}}{n} \sum_{k=1}^u \frac{4n}{c^2 k^2} + \sum_{k>u} \frac{e^{-ck/2n^{\frac{1}{2}}}}{(1 - e^{-ck/2n^{\frac{1}{2}}})^2} + O\left(\frac{e^{cn^{\frac{1}{2}}}}{n^{\frac{1}{2}+\epsilon}}\right), \quad u = [n^{\frac{1}{2}}].$$

But

$$\sum_{k=1}^u \frac{4n}{c^2 k^2} = \frac{4n}{c^2} \frac{\pi^2}{6} - \frac{4n}{c^2} \sum_{k>u} \frac{1}{k^2} = n - \frac{4n}{c^2 u} + O\left(\frac{n}{u^2}\right).$$

And

$$\begin{aligned}\sum_{k>u} \frac{e^{-ck/2n^{\frac{1}{2}}}}{(1 - e^{-ck/2n^{\frac{1}{2}}})^2} &= \int_u^{\infty} \frac{e^{-cx/2n^{\frac{1}{2}}}}{(1 - e^{-cx/2n^{\frac{1}{2}}})^2} dx + O\left(\frac{1}{u^2}\right) \\ &= \frac{2n^{\frac{1}{2}}}{c(1 - e^{-cu/2n^{\frac{1}{2}}})} - \frac{n^{\frac{1}{2}}}{c} + O\left(\frac{1}{u^2}\right) = \frac{4n}{c^2 u} - \frac{n^{\frac{1}{2}}}{c} + O\left(\frac{1}{n^{\frac{1}{2}+\epsilon}}\right).\end{aligned}$$

Thus finally

$$\Sigma'_1 = e^{cn^{\frac{1}{2}}} - \frac{e^{cn^{\frac{1}{2}}}}{cn^{\frac{1}{2}}} + O\left(\frac{e^{cn^{\frac{1}{2}}}}{n^{\frac{1}{2}+\epsilon}}\right).$$

Hence

$$\Sigma = \Sigma'_1 - \Sigma''_1 + \Sigma_2 = e^{cn^{\frac{1}{2}}} \left[1 + O\left(\frac{1}{n^{\frac{1}{2}+\epsilon}}\right) \right]$$

which proves the lemma.

Next we show that

$$(7) \quad 0 < \liminf \frac{np(n)}{e^{cn^{\frac{1}{2}}}} \leq \limsup \frac{np(n)}{e^{cn^{\frac{1}{2}}}} < \infty.$$

To prove (7) write

$$(8) \quad c_1^{(n)} = \max_{m \leq n} \frac{mp(m)}{e^{cm^{\frac{1}{2}}}}.$$

Clearly by (8) and (6) and (2)

$$(n+1)p(n+1) \leq c_1^{(n)} \sum_{v=1}^n \sum_{\substack{k=1 \\ kv < n}} \frac{ve^{c(n+1-kv)^{\frac{1}{2}}}}{n+1-kv} < c_1^{(n)} e^{c(n+1)^{\frac{1}{2}}} \left(1 + \frac{b_1}{n^{\frac{1}{2}+\epsilon}}\right)^4.$$

Write

$$\frac{(n+j)p(n+j)}{e^{c(n+j)^{\frac{1}{2}}}} = c_1^{(n)} \left(1 + \frac{b_j}{n^{\frac{1}{2}+\epsilon}}\right), \quad j = 1, 2,$$

Then

$$\begin{aligned} (n+r+1)p(n+r+1) &< c_1^{(n)} \sum_{v=1}^n \sum_{\substack{k=1 \\ kv \leq n+r}} \frac{ve^{c(n+r+1-kv)^{\frac{1}{2}}}}{n+r+1-kv} \\ &\quad + c_1^{(n)} \frac{\max_{i \leq r} b_i}{n^{\frac{1}{2}+\epsilon}} \sum_{v=1}^n \sum_{\substack{k=1 \\ kv \leq r}} \frac{ve^{c(n+r+1-kv)^{\frac{1}{2}}}}{n+r+1-kv} \\ &< c_1^{(n)} e^{c(n+r+1)^{\frac{1}{2}}} \left(1 + \frac{b_1}{n^{\frac{1}{2}+\epsilon}} + \frac{\max_{i \leq r} b_i}{n^{\frac{1}{2}+\epsilon}} \frac{r^2 e^{c(n+r+1)^{\frac{1}{2}}}}{n}\right), \end{aligned}$$

since

$$\sum_{kv \leq r} v \leq r^2.$$

Hence

$$b_{r+1} < b_1 + \frac{r^2 \max_{i \leq r} b_i}{n}.$$

We show that, for $r^2 \leq n/2$, $b_{r+1} \leq 2b_1$. We use induction. We have

$$b_{r+1} < b_1 + \frac{r^2 \cdot 2b_1}{n} \leq 2b_1.$$

* b_1 is chosen such that for every $m > 0$

$$\sum_v \sum_{\substack{k \\ m-kv > 0}} \frac{ve^{c(m-kv)^{\frac{1}{2}}}}{m-kv} < e^{c(m)^{\frac{1}{2}}} \left(1 + \frac{b_1}{m^{\frac{1}{2}+\epsilon}}\right).$$

Thus

$$c_1^{[n+(\frac{1}{2}n)^{\frac{1}{2}}]} \leq c_1^{(n)} \left(1 + \frac{2b_1}{n^{\frac{1}{2}+\epsilon}} \right).$$

Or

$$c_1^{((m+1)^2)} < c_1^{(m^2)} \left(1 + \frac{10b_1}{n^{\frac{1}{2}+\epsilon}} \right);$$

and since $\sum m^{1/1+\epsilon}$ converges we see that $\limsup c_1^{(n)} < \infty$; i.e. $\limsup np(n)/e^{cn^{\frac{1}{2}}} < \infty$. Similarly we can show that $\liminf np(n)/e^{cn^{\frac{1}{2}}} > 0$, which completes the proof of (7).

Next we prove that

$$(9) \quad \liminf \frac{np(n)}{e^{cn^{\frac{1}{2}}}} = \limsup \frac{np(n)}{e^{cn^{\frac{1}{2}}}}$$

and this will complete the proof of (1).

Suppose that (9) does not hold; write

$$(10) \quad \liminf \frac{np(n)}{e^{cn^{\frac{1}{2}}}} = d, \quad \limsup \frac{np(n)}{e^{cn^{\frac{1}{2}}}} = D.$$

Now choose n large and such that

$$\frac{np(n)}{e^{cn^{\frac{1}{2}}}} > D - \epsilon.$$

Then since $p(n)$ is an increasing function of n there exists a c_2 such that for every m in the range $n \leq m \leq n + c_2 n^{\frac{1}{2}}$

$$\frac{mp(m)}{e^{cm^{\frac{1}{2}}}} > \frac{d + D}{2}.$$

Now we claim that for every r_1 there exists a $\delta_{r_1} = \delta(r_1)$ such that, for $n \leq m \leq n + r_1 n^{\frac{1}{2}}$,

$$(11) \quad \frac{mp(m)}{e^{cm^{\frac{1}{2}}}} > d + \delta_{r_1}.$$

We prove (11) as follows: We evidently have by our lemma

$$mp(m) \geq d \sum_{v=1}^m \sum_{\substack{k=1 \\ kv < m}} \frac{ve^{c(m-kv)^{\frac{1}{2}}}}{m - kv} + \frac{D - d}{2} \sum_{\substack{v=1 \\ n \leq m-kv \leq n+c_2 n^{\frac{1}{2}}}} \sum_{k=1} \frac{ve^{c(m-kv)^{\frac{1}{2}}}}{m - kv} - o(e^{cm^{\frac{1}{2}}})^5$$

⁵ The term $o(e^{cm^{\frac{1}{2}}})$ is present because d is the lower limit and not the lower bound of $\frac{mp(m)}{e^{cm^{\frac{1}{2}}}}$.

$$\begin{aligned}
&> de^{cm^{\frac{1}{2}}} + \frac{D-d}{2} \frac{e^{cn^{\frac{1}{2}}}}{m} \sum_{n \leq m-v \leq n+c_2 n^{\frac{1}{2}}} v - o(e^{cm^{\frac{1}{2}}}) > de^{cm^{\frac{1}{2}}} + c_3 e^{cn^{\frac{1}{2}}} - o(e^{cm^{\frac{1}{2}}}) \\
&> (d + \delta_{r_1}) e^{cm^{\frac{1}{2}}}, \quad \left(\text{i.e. } \frac{e^{cn^{\frac{1}{2}}}}{e^{cm^{\frac{1}{2}}}} > c_4 \right).
\end{aligned}$$

which proves (11).

Suppose $2n \geq m \geq n + sn^{\frac{1}{2}}$, s sufficiently large; we show that

$$(12) \quad \sum_{\substack{v=1 \\ m-kv < n}} \sum_{k=1} v \frac{e^{c(m-kv)^{\frac{1}{2}}}}{m-kv} < \frac{e^{cm^{\frac{1}{2}}}}{s^{10}}.$$

Clearly

$$\begin{aligned}
\sum_{\substack{v \\ 0 < m-kv < n}} \sum_k \frac{ve^{c(m-kv)^{\frac{1}{2}}}}{m-kv} &\leq \sum_{\substack{v \\ kv > sn^{\frac{1}{2}}}} \sum_k \frac{ve^{c(m-kv)^{\frac{1}{2}}}}{m-kv} \\
&< e^{cm^{\frac{1}{2}}} \sum_{\substack{v \\ \frac{1}{2}m > kv > sn^{\frac{1}{2}}}} \sum_k \frac{2ve^{-ckv/2m^{\frac{1}{2}}}}{m} + \sum_{m > kv \geq \frac{1}{2}m} \sum_k \frac{ve^{c(m-kv)^{\frac{1}{2}}}}{m-kv} \\
&< e^{cm^{\frac{1}{2}}} \sum_{\substack{v \\ \frac{1}{2}m > kv > sn^{\frac{1}{2}}}} \sum_k \frac{2ve^{-ckv/2m^{\frac{1}{2}}}}{m} + m^2 e^{c(\frac{1}{2}m)^{\frac{1}{2}}},
\end{aligned}$$

since

$$\sum_{\substack{v \\ kv < x}} \sum_k v \leq x^2.$$

Further

$$\begin{aligned}
\sum_{\substack{v \\ \frac{1}{2}m > kv > sn^{\frac{1}{2}}}} \sum_k ve^{-ckv/2m^{\frac{1}{2}}} &< \sum_{u=1}^m \sum_{\substack{v \\ (u+1)sn^{\frac{1}{2}} \geq kv > usn^{\frac{1}{2}}}} \sum_k ve^{-cusn^{\frac{1}{2}}/2m^{\frac{1}{2}}} \\
&< \sum_{u=1}^m \sum_{\substack{v=1 \\ kv \leq (u+1)sn^{\frac{1}{2}}}} \sum_{k=1} ve^{-cuu/4} < \sum_{u=1}^m (u+1)^2 s^2 ne^{-cus/4}.
\end{aligned}$$

Thus

$$\sum_{\substack{v \\ \frac{1}{2}m > kv > sn^{\frac{1}{2}}}} \sum_k ve^{-ckv/2m^{\frac{1}{2}}} < ms^2 \sum_{u=1}^{\infty} (u+1)^2 e^{-cus/4} < \frac{m}{4s^{10}}$$

for sufficiently large s . Hence finally

$$\sum_{\substack{v=1 \\ m-kv < n}} \sum_{k=1} \frac{ve^{c(m-kv)^{\frac{1}{2}}}}{m-kv} < \frac{e^{cm^{\frac{1}{2}}}}{2s^{10}} + m^2 e^{c(\frac{1}{2}m)^{\frac{1}{2}}} < \frac{e^{cm^{\frac{1}{2}}}}{s^{10}}$$

for sufficiently large m and s (since $s < n^{\frac{1}{2}}$).

Consider now the intervals $n + tn^{\frac{1}{t}}, n + (t+1)n^{\frac{1}{t}}, t > r_1, t+1 < n^{\frac{1}{t}}$. Split it into t^2 equal parts. Write

$$\min \frac{mp(m)}{e^{cm^{\frac{1}{t}}}} = d + \delta_t^u, \quad n \leq m \leq n + \left(t + \frac{u+1}{t^2}\right)n^{\frac{1}{t}}$$

and put $\delta_t^{t^2-1} = \delta_t$. Now let $n + (t + u/t^2)n^{\frac{1}{t}} \leq m \leq n + (t + (u+1)/t^2)n^{\frac{1}{t}}$; then we have

$$mp(m) > d \sum_{v=1} \sum_{\substack{k=1 \\ kv < m}} \frac{ve^{c(m-kv)^{\frac{1}{t}}}}{m - kv} + \delta_t^{(u-1)} \sum'_v \sum'_k \frac{ve^{c(m-kv)^{\frac{1}{t}}}}{m - kv} - o(e^{cm^{\frac{1}{t}}}),$$

where the primes indicate that the summation is extended only over those v and k for which $n \leq m - kv \leq n + (t + u/t^2)n^{\frac{1}{t}}$. Further by Lemma 1

$$\begin{aligned} mp(m) \geq (d + \delta_t^{(u-1)})e^{cm^{\frac{1}{t}}} - \delta_t^{(u-1)} \sum'' \frac{ve^{c(m-kv)^{\frac{1}{t}}}}{m - kv} \\ - \delta_t^{(u-1)} \sum''' \frac{ve^{c(m-kv)^{\frac{1}{t}}}}{m - kv} - o(e^{cm^{\frac{1}{t}}}), \end{aligned}$$

where in \sum'' the summation is extended only over those v and k for which $m - kv \leq n$, and in \sum''' the summation is extended only over those v and k for which $m - kv \geq n + (t + u/t^2)n^{\frac{1}{t}}$. We have by (11)

$$\sum'' < \frac{e^{cm^{\frac{1}{t}}}}{t^{10}}.$$

Further we have

$$\sum''' < \frac{n}{t^4} \frac{2e^{cm^{\frac{1}{t}}}}{m} < \frac{2e^{cm^{\frac{1}{t}}}}{t^4}.$$

Hence finally

$$mp(m) > e^{cm^{\frac{1}{t}}} \left(d + \delta_t^{(u-1)} - \frac{3\delta_t^{(u-1)}}{t^4} \right) - o(e^{cm^{\frac{1}{t}}}).$$

Hence

$$\delta_t^{(u)} > \delta_t^{(u-1)} \left(1 - \frac{3}{t^4} \right) - o(1).$$

Thus if t is fixed, independent of n , we have

$$\delta_{t+1} > \delta_t \left(1 - \frac{3}{t^4} \right)^{t^2} - o(1),$$

therefore

$$\delta_t > \delta_{r_1} \prod_{u > r_1} \left(1 - \frac{3}{u^4} \right)^{u^2} - o(1).$$

But $\prod_u (1 - 3/u^4)^{u^2}$ converges; thus, if r_1 was sufficiently large, we have $\delta_t > \delta_{r_1}/2$. Now choose r_2 sufficiently large; then we have $\delta_{r_2} > \delta_{r_1}/2$, i.e. for $n \leq m \leq n + r_2 n^{\frac{1}{2}}$,

$$\frac{mp(m)}{e^{cm^{\frac{1}{2}}}} > d + \frac{\delta_{r_1}}{2}.$$

Consider the interval $n + tn^{\frac{1}{2}}, n + (t+1)n^{\frac{1}{2}}, t > r_2$. Split it into t^2 equal parts. $\delta_t^{(u)}$ and δ_t have the same meaning as before. Suppose $n + (t + u/t^2)n^{\frac{1}{2}} \leq m \leq n + (t + (u+1)/t^2)n^{\frac{1}{2}}$; then evidently

$$mp(m) > (d + \delta_t^{(u-1)}) \sum'_v \sum'_k \frac{ve^{c(m-kv)^{\frac{1}{2}}}}{m - kv},$$

where the primes indicate that the summation is extended only over those v and k for which $n \leq m - kv \leq n + n^{\frac{1}{2}}(t + u/t^2)$.

Now

$$\sum'_v \sum'_k \frac{ve^{c(m-kv)^{\frac{1}{2}}}}{m - kv} = \sum_{v=1} \sum_{\substack{k=1 \\ kv < m}} \frac{ve^{c(m-kv)^{\frac{1}{2}}}}{m - kv} = \sum'' + \sum''',$$

where \sum'' and \sum''' are defined as before. By (12) and the previous estimate of \sum''' we have

$$\sum'' < \frac{e^{cm^{\frac{1}{2}}}}{t^{10}}, \quad \sum''' < \frac{2e^{cm^{\frac{1}{2}}}}{t^4}.$$

Hence by Lemma 1

$$mp(m) > e^{cm^{\frac{1}{2}}}(d + \delta_t^{(u-1)}) \left(1 - \frac{3}{t^4}\right) - \frac{b_1(d + \delta_t^{(u-1)})e^{cm^{\frac{1}{2}}}}{n^{\frac{1}{2}+s}};$$

i.e.

$$d + \delta_t^{(u)} > (d + \delta_t^{(u-1)}) \left(1 - \frac{3}{t^4}\right) - \frac{b_1(d + \delta_t^{(u-1)})}{n^{\frac{1}{2}+s}},$$

and

$$d + \delta_{t+1} > (d + \delta_t) \left(1 - \frac{3}{t^4}\right)^{t^2} - \frac{b_1 t^2 (d + \delta_t^{(u-1)})}{n^{\frac{1}{2}+s}},$$

or

$$d + \delta_s > \left(d + \frac{\delta_{r_1}}{2}\right) \prod_{t > r_2} \left(1 - \frac{3}{t^4}\right)^{t^2} - \frac{b_2 s^3}{n^{\frac{1}{2}+s}}.$$

For sufficiently large r_2 we have,

$$\left(d + \frac{\delta_{r_1}}{2}\right) \prod_{t > r_2} \left(1 - \frac{3}{t^4}\right)^{t^2} > d + \frac{\delta_{r_1}}{4},$$

and if $s \leq (\log n)^2$ and n is sufficiently large,

$$\delta_s > \frac{\delta_{r_1}}{8};$$

that is, for $n \leq m \leq n + n^{\frac{1}{2}}(\log n)^2$

$$\frac{mp(m)}{e^{cm^{\frac{1}{2}}}} > d + \frac{\delta_{r_1}}{8}.$$

Now suppose $m > n + n^{\frac{1}{2}}(\log n)^2$; we shall show that

$$\sum = \sum_v \sum_k \frac{ve^{c(m-kv)^{\frac{1}{2}}}}{m - kv} < \frac{e^{cm^{\frac{1}{2}}}}{m}.$$

We have

$$\sum < m^2 e^{c\gamma^{\frac{1}{2}}} < m^2 e^{cm^{\frac{1}{2}} - 10c \log m} < \frac{e^{cm^{\frac{1}{2}}}}{m}$$

for sufficiently large n . Hence by Lemma 1,

$$\sum_v \sum_k \frac{ve^{c(m-kv)^{\frac{1}{2}}}}{m - kv} > e^{cm^{\frac{1}{2}}} \left(1 - \frac{b'_1}{n^{\frac{1}{2}+\epsilon}}\right).^6$$

Now we continue as in the proof of (7). Suppose $t > n + n^{\frac{1}{2}}(\log n)^2$; write

$$d + \delta_t = \min \frac{mp(m)}{e^{cm^{\frac{1}{2}}}}, \quad n \leq m \leq t.$$

Then

$$(t+1)p(t+1) \geq (d + \delta_t) \sum_v \sum_k \frac{ve^{c(t-kv)^{\frac{1}{2}}}}{t - kv} > (d + \delta_t)e^{ct^{\frac{1}{2}}} \left(1 - \frac{b'_1}{t^{\frac{1}{2}+\epsilon}}\right).$$

Write

$$\frac{(t+r)p(t+r)}{e^{c(t+r)^{\frac{1}{2}}}} = (d + \delta_t) \left(1 - \frac{b'_r}{t^{\frac{1}{2}+\epsilon}}\right).$$

Then as in the proof of (7) we have

$$\begin{aligned} (t+j+1)p(t+j+1) &> (d + \delta_t) \sum_v \sum_k \frac{ve^{c(t+j+1-kv)^{\frac{1}{2}}}}{t+j+1 - kv} \\ &\quad - (d + \delta_t) \frac{\max_{r \leq j} b'_r}{t^{\frac{1}{2}+\epsilon}} \frac{j^2}{t} e^{c(t+j+1)^{\frac{1}{2}}} \\ &> (d + \delta_t)e^{c(t+j+1)^{\frac{1}{2}}} \left(1 - \frac{b'_1}{(t+j+1)^{\frac{1}{2}}}\right) - (d + \delta_t) \frac{\max_{r \leq j} b'_r}{t^{\frac{1}{2}+\epsilon}} \frac{j^2}{t} e^{c(t+j+1)^{\frac{1}{2}}} \\ &= (d + \delta_t)e^{c(t+j+1)^{\frac{1}{2}}} \left(1 - \frac{b'_{j+1}}{t^{\frac{1}{2}+\epsilon}}\right), \end{aligned}$$

⁶ As in footnote 4 b'_1 is chosen such that for every $m > n + n^{\frac{1}{2}}(\log n)^2$

$$\sum_v \sum_k \frac{ve^{c(m-kv)^{\frac{1}{2}}}}{m - kv} > e^{cm^{\frac{1}{2}}} \left(1 - \frac{b'_1}{m^{\frac{1}{2}+\epsilon}}\right).$$

where

$$b'_{j+1} < b'_1 + \max_{r \leq j} b'_r \cdot \frac{j^2}{t}.$$

We show that for $j^2 < t/2$ we have, $b'_{j+1} < 2b'_1$. We use induction; we have

$$b'_{j+1} < b'_1 + \frac{2b'_1}{2} = 2b'_1.$$

Thus

$$d + \delta_{[t+\frac{1}{2}t]} > (d + \delta_t) \left(1 - \frac{2b'_1}{t^{1+\epsilon}}\right).$$

That is,

$$d + \delta_{(s+1)^2} > (d + \delta_{s^2}) \left(1 - \frac{10b'_1}{s^{1+\epsilon}}\right).$$

Therefore

$$d + \delta_{u^2} > \left(d + \frac{\delta_{r_1}}{8}\right) \prod_{v > \log n} \left(1 - \frac{10b'_1}{v^{1+\epsilon}}\right) > d + \frac{\delta_{r_1}}{10},$$

which contradicts (10); and this completes the proof of (1).

As can be seen, the main idea of our proof is rather simple; unfortunately the details are long and cumbersome. By the same method we can prove the following result: Let m be a fixed integer. Denote by $p_{a_1, a_2, \dots, a_r}^{(m)}(n)$ the number of partitions of n into integers congruent to one of the numbers $a_1, a_2, \dots, a_r \pmod{m}$. Then

$$(13) \quad p_{a_1, a_2, \dots, a_r}^{(m)}(n) \sim \frac{a}{n^\alpha} e^{cn^{\frac{1}{2}}}, \quad ((a_1, a_2, \dots, a_r)^m = 1)$$

where C depends on m and r , and α and a depend on m, a_1, a_2, \dots, a_r .

The same method will work if we consider partitions of n into r th powers. Denote the number of partitions of n into r th powers by $p_r(n)$, Hardy, Ramanujan and Wright⁷ proved that

$$(14) \quad p_r^{(n)} \sim c_1 n^{\frac{1}{(r+1)-\frac{1}{2}}} e^{c_2 n^{1/(r+1)}}.$$

Clearly as in the case of $p(n)$ we have

$$np_r(n) = \sum_{\substack{v \\ vk < n}} \sum_k v^r p_r(n - kv^r).$$

⁷ Hardy, Ramanujan, *ibid.* p. 111. Maitland Wright, *Acta Math.* 63, (1934), pp. 143–191. Wright proves a very much sharper result than (13).

To prove (14) we should only have to prove the analogue of our lemma, namely

$$(15) \quad \sum_{\substack{v \\ v^r k < n}} \sum_k (n - v^r k)^{1/(r+1)-\frac{1}{2}} e^{c_2(n-v^r k)^{1/(r+1)}} \\ = n^{1/(r+1)-\frac{1}{2}} e^{c_2 n^{1/(r+1)}} \left[1 + O\left(\frac{1}{n^{1-(1/(r+1))+\epsilon}}\right) \right].$$

If (15) is proved the proof of (14) proceeds as in the case of $p(n)$.

I have not worked out a proof of (15); it seems likely that a proof would be longer than that of Lemma 1, but would not present any particular difficulties.

Recently Ingham⁸ proved a Tauberian theorem from which (1) and (14) follow as corollaries. In fact his Theorem 2 gives a more general result, from which (13) also follows as a very special case.

Denote by $P_r(n)$ the number of partitions of n into powers of r . Clearly

$$nP_r(n) = \sum_{\substack{v \\ r^v k < n}} \sum_k r^v P_r(n - r^v k).$$

It might be possible to get an asymptotic formula for $P_r(n)$ by our method. I have not succeeded so far. But we can show without difficulty that

$$(16) \quad \log P_r(n) \sim \frac{(\log n)^2}{2 \log r}.$$

We have

$$f(x) = \sum_{n=0}^{\infty} P_r(n) x^n = \prod_{r=1}^{\infty} \frac{1}{1 - x^{r^i}}.$$

It is easy to see that for $0 \leq x \leq 1$,

$$(17) \quad c_1 \left(\frac{1}{1-x} \right)^{(1/(2 \log a)) \log 1/(1-x)} < f(x) < c_2 \left(\frac{1}{1-x} \right)^{(1/(2 \log a)) \log 1/(1-x)}.$$

Thus

$$P_r(n) \left(1 - \frac{1}{n} \right)^n < f \left(1 - \frac{1}{n} \right) < c_2 n^{(\log n)/(2 \log a)};$$

that is

$$P_r(n) < c_3 n^{(\log n)/(2 \log a)}, \quad \log P_r(n) < (1 - \epsilon) \frac{(\log n)^2}{2 \log a} \quad \text{for } n > n_0.$$

Suppose now that for a certain large n $\log(P_r(n)) < (1 - \epsilon)(\log n)^2/2 \log a$; then, since for $m < n$ $P_r(m) \leq P_r(n)$ we have

$$(18) \quad f(x) < e^{(1-\epsilon) \cdot (\log n)^2/(2 \log a)} \sum_{k=0}^n x^k + \sum_{k>n} c_3 k^{(\log k)/(2 \log a)} x^k,$$

⁸ A. E. Ingham, *A Tauberian Theorem for Partitions*, these Annals, 42 (1941), p. 1083.

and a simple calculation shows that (18) contradicts (17). (Choose $x = (1 - \delta)n$, $\delta = \delta(\epsilon)$). The same method would of course give

$$\log(p(n)) \sim \pi \left(\frac{2n}{3}\right)^{\frac{1}{2}}.$$

We can also prove the following results:

I. Let $a_1 < a_2 < \dots$ be an infinite sequence of integers of density α , such that the a 's have no common factor. Denote by $p'(n)$ the number of partitions of n into the a 's. Then

$$(19) \quad \log(p'(n)) \sim c(\alpha n)^{\frac{1}{2}}. \quad (c = \pi(\frac{2}{3})^{\frac{1}{2}})$$

II. Let $a_1 < a_2 < \dots$ be an infinite sequence of integers of density α , such that every sufficiently large m can be expressed as the sum of different a 's. Then denote by $P'(n)$ the number of partitions of n into different a 's. Then

$$(20) \quad \log P'(n) \sim c \left(\frac{\alpha}{2} n\right)^{\frac{1}{2}}.$$

We shall sketch the proof of II; the proof of I is similar but simpler. Denote by $P(n)$ the number of partitions of n into different summands: it is well known that⁹

$$(21) \quad \log P(n) \sim c \left(\frac{n}{2}\right)^{\frac{1}{2}}.$$

First we show that

$$(22) \quad \limsup \frac{\log P'(n)}{c \left(\frac{\alpha}{2} n\right)^{\frac{1}{2}}} \leq 1.$$

To the partition $n = a_{i_1} + a_{i_2} + \dots + a_{i_r}$ we make correspond the partition $i_1 + i_2 + \dots + i_r$. For $i > i_0$ we have $i < a_i(\alpha + \epsilon)$ therefore $i_1 + i_2 + \dots + i_k < n(\alpha + \epsilon) + i_0^2$. Thus each partition of n into the a 's is mapped into a partition of integers $\leq n(\alpha + 2\epsilon)$ with all integers as summands; hence from (20) we obtain (22). Next we prove that

$$(23) \quad \liminf \frac{\log P'(n)}{c \left(\frac{\alpha}{2} n\right)^{\frac{1}{2}}} \geq 1.$$

Split the sequence a_i into two disjoint sequences b_1, b_2, \dots and c_1, c_2, \dots where the b 's have density 0 and every sufficiently large integer is the sum of different b 's and the c 's are the remaining a 's. It is easy to see that we can find the b 's with the required property; also the density of the c 's is clearly α . Denote by $Q(n)$ the number of partitions of n into the c 's. Now associate

⁹ A well known result of Euler states that the number of partitions of n into odd integers equals the number of partitions of n into different summands. Thus (20) follows from i.

with the partition $n = i_1 + i_2 + \cdots + i_k$, $i_1 < i_2 < \cdots < i_k$ the partition $c_{i_1} + c_{i_2} + \cdots + c_{i_k}$; as before, we have

$$\frac{n}{\alpha + \epsilon} < c_{i_1} + c_{i_2} + \cdots + c_{i_k} < \frac{n}{\alpha - \epsilon}.$$

Hence for at least one $n/(\alpha + \epsilon) < m < n/(\alpha - \epsilon)$, $Q(m) > p(n)(\alpha - \epsilon)/n$. Thus there exists a sequence of integers $x_1 < x_2 < \cdots$ with $\lim x_{i+1}/x_i = 1$ and

$$(24) \quad \liminf \frac{\log Q(x_i)}{c \left(\frac{\alpha}{2} x_i \right)^{\frac{1}{\alpha}}} = 1.$$

Now suppose $x_j \leq m < x_{j+1}$. Choose x_i such that $\epsilon m > m - x_i > C$. Then $m - x_i$ is a sum of different b 's, hence $P(m) \geq Q(x_i)$. Thus (23) follows from (24), and this completes the proof of II.

If might be worth while to mention the following problem: Let $a_1 < a_2 < \cdots$ be an infinite sequence of integers, such that $\log P(n) \sim c(\alpha n)^{\frac{1}{\alpha}}$. Does it then follow that the density of the a 's is α . I cannot decide this problem. Perhaps the following result might be of some interest in this connection: Let $a_1 < a_2 < \cdots$ be an infinite sequence of integers. $f(n)$ denotes the number of a 's $\leq n$, and $\varphi(n)$ denotes the number of solutions of $a_i + a_j \leq n$. It can be shown trivially that if $\lim f(n)/n^\alpha = c_1$ then $\lim \varphi(n)/n^{2\alpha} = c_2$. But the converse is also true, and can be simply proved by using a Tauberian theorem of Hardy and Littlewood.¹⁰ We have

$$(f(z))^2 = \left(\sum_{i=1}^{\infty} z^{a_i} \right)^2 = \sum_{k=1}^{\infty} d_k z^k$$

and, since $\sum d_k = \varphi(n) \sim c_2 n^{2\alpha}$, we evidently have

$$f(z) \sim_{z \rightarrow 1} \frac{c_3}{(1-z)^\alpha}$$

and hence by the theorem of Hardy and Littlewood,

$$f(n) = \sum_{a_k \leq n} 1 \sim c_1 n^\alpha.$$

By the same methods that were used in proving II, we can prove the following result: Denote by $R(n)$ the number of partitions of n into integers relatively prime to n . We have

$$\log R(n) \sim c(\varphi(n))^{\frac{1}{\alpha}}.$$

Similarly, if we denote by $R'(n)$ the number of partitions of n into different integers relatively prime to n , we have

$$\log R'(n) \sim c \left(\frac{\varphi(n)}{2} \right)^{\frac{1}{\alpha}}.$$

¹⁰ Hardy-Littlewood, *Tauberian Theorems*, Proc. London Math. Soc. 13, (1914), pp. 174-191.

I have not succeeded in finding asymptotic formulas for $R(n)$ and $R'(n)$. This problem seems rather difficult.

March 12, 1942.

In the meantime I have proved the above conjecture. Consider

$$f(x) = \sum_{n=1}^{\infty} P(n)x^n = \prod_{k=1}^{\infty} \frac{1}{1-x^{a_k}}.$$

If we assume that $\log P(n) \sim a(n)^{\frac{1}{2}}$, we obtain by a simple calculation

$$\log f(x) \underset{x \rightarrow 1}{\sim} \frac{\pi^2}{6} \frac{\alpha}{1-x}.$$

But

$$\log f(x) = \sum x^{a_i} + \frac{1}{2} \sum x^{2a_i} + \dots = \sum_{k=1}^{\infty} b_k x^k.$$

Denote by $A(n)$ the number of a 's not exceeding n . We have

$$B(n) = \sum_{k=1}^n b_k = \sum_{k=1}^{\infty} \frac{1}{k} A\left(\frac{n}{k}\right).$$

Thus

$$A(n) = \sum_{k=1}^{\infty} \frac{u(k)}{k} B\left(\frac{n}{k}\right).$$

But by the well known Tauberian theorem of Hardy-Littlewood,¹¹ we have

$$B(n) \sim \frac{\alpha \pi^2 n}{6}.$$

Hence

$$A(n) \sim \sum_{k=1}^{\infty} \frac{u(k)}{k^2} \cdot \frac{\alpha \pi^2 n}{6} \sim \alpha n. \quad \text{q.e.d.}$$

Similarly we can show that if $\log P'(n) = c[(\alpha/2)n]^{\frac{1}{2}}$, the density of the a 's is α .

UNIVERSITY OF PENNSYLVANIA

¹¹ Hardy-Littlewood, *ibid.*

ON A PROBLEM OF I. SCHUR

BY P. ERDÖS AND G. SZEGÖ

(Received June 11, 1941)

TO THE MEMORY OF I. SCHUR

1. Introduction

1. Let $n \geq 3$, and let Q_n denote the class of polynomials $f(x)$ of degree n satisfying the condition $|f(x)| \leq 1$ in the interval $-1 \leq x \leq +1$. Let $Q_n(x_0)$ denote the subclass of Q_n characterized by the further restriction $f''(x_0) = 0$.

A well-known theorem of A. Markoff¹ states that $|f'(x)| \leq n^2$ for $-1 \leq x \leq +1$ provided that $f(x) \in Q_n$; here $|f'(x)| = n^2$ holds if and only if $x = \pm 1$ and $f(x) = \pm T_n(x)$, where $T_n(x)$ denotes the n^{th} Tchebycheff polynomial. We observe that $T_n(x)$ does not belong to the classes $Q_n(\pm 1)$.

Some years ago I. Schur² proved the following interesting theorem: *Let $-1 \leq x_0 \leq +1$, and let $f(x)$ belong to $Q_n(x_0)$. Then $|f'(x_0)| < \frac{1}{2}n^2$. Moreover he showed: Let m_n be the least positive constant (depending only on n) such that $|f'(x_0)| \leq m_n \cdot n^2$ for all $f(x) \in Q_n(x_0)$, and x_0 in $-1 \leq x \leq +1$. If $\mu = \limsup_{n \rightarrow \infty} m_n$, then*

$$(1.1) \quad 0.217 \cdots \leq \mu \leq 0.465 \cdots$$

Obviously

$$(1.2) \quad m_n \cdot n^2 = \max_{-1 \leq x_0 \leq +1} \max_{f(x) \in Q_n(x_0)} |f'(x_0)|.$$

The main purpose of the present note is to determine the constant μ and the polynomial $f(x)$ for which the extremum (1.2) is attained. In terms of the constant m_n , we obtain a bound for the derivative $f'(x)$ of a polynomial $f(x)$ which satisfies the condition that $|f'(x)|$ has a relative maximum at the point x considered.

2. Let $u_n(x)$ be the polynomial of the class $Q_n(+1)$ for which $u'_n(1)$ is a maximum. This polynomial $u_n(x) = u_n(x; A_n)$ can be determined from the transcendental equations (2.5), (2.6) and (2.17) of §2 (see below). It is a special case of a remarkable class of polynomials $u_n(x; A)$ considered first by G. Zolotareff³

¹ A. Markoff, *On a certain problem of D. I. Mendeleeff* (in Russian), *Zapiski Imperatorskoi Akademii Nauk*, vol. 62 (1889), pp. 1-24.

² I. Schur, *Über das Maximum des absoluten Betrages eines Polynoms in einem gegebenen Intervall*, *Mathematische Zeitschrift*, vol. 4 (1919), pp. 271-287.

³ G. Zolotareff, (a) *On a question concerning a minimum value* (in Russian), Dissertation "pro venia legendi," published in lithographed form, 1868, *Oeuvres*, vol. 2 (1902), pp. 130-166; (b) *Application of elliptic functions to questions concerning functions which deviate the least from zero* (in Russian), *Zapiski Imperatorskoi Akademii Nauk*, vol. 30 (1877), *Oeuvres*, vol. 2, pp. 1-59; (c) *Sur l'application des fonctions elliptiques aux questions de maxima et minima*, *Bulletin de l'Académie des Sciences de St.-Petersbourg*, series 3, vol. 24 (1878), pp. 305-310, *Mélanges*, 5, pp. 419-426, *Oeuvres*, vol. 1 (1931), pp. 369-374.

playing also a role in the important investigations of W. Markoff.⁴ Recently N. Achyesser⁵ used polynomials of the Zolotareff type in his investigations on polynomials of least deviation in two disjoint intervals. With the previous notation, our main result is:

THEOREM 1. *The extremum $m_n \cdot n^2$ in (1.2) is attained for $x_0 = +1$ and for the Zolotareff polynomials $\pm u_n(x)$ [or for $x_0 = -1$ and for $\pm u_n(-x)$], provided n is sufficiently large. Furthermore*

$$(1.3) \quad \lim_{n \rightarrow \infty} m_n = \mu$$

exists and

$$(1.4) \quad \mu = k^{-2}(1 - E/K)^2 = 0.3124 \dots,$$

where k^2 is the only root of the transcendental equation

$$(1.5) \quad (K - E)^3 + (1 - k^2)K - (1 + k^2)E = 0$$

satisfying the condition $0 < k^2 < 1$. Here K and E are the complete elliptic integrals associated with the modulus k .

A further analysis and discussion of a few special cases furnishes the more informative

THEOREM 2. *If $n > 3$ the extremum $m_n \cdot n^2$ in (1.2) is attained in the cases mentioned in Theorem 1, and only in these cases. If $n = 3$, it is attained for $x_0 = 0$ and for the Tchebycheff polynomials $\pm T_3(x)$, and only then.*

In §§2 and 3 of the present paper we first study as a preparation the general polynomials $u_n(x; A)$ of Zolotareff and the special case $u_n(x) = u_n(x; A_n)$ mentioned above. The proof of Theorem 1 is then given in §§4 and 5, and that of Theorem 2 in §§6 and 7. In §8 we consider two problems of Zolotareff in which the polynomials $u_n(x; A)$ were first used; §9 contains another application.

The polynomials of Zolotareff occur in numerous other related extremum problems. They satisfy a simple differential equation by means of which they can be brought in relationship with the multiplication problem of elliptic integrals. In what follows we have tried to reduce the use of elliptic functions to a minimum.⁶

⁴ W. Markoff, *Über Polynome, die in einem gegebenen Intervalle möglichst wenig von Null abweichen*, Mathematische Annalen, vol. 77 (1916), pp. 213-258. The Russian original appeared 1892.

⁵ N. Achyesser, (a) *Über einige Funktionen, welche in zwei gegebenen Intervallen am wenigsten von Null abweichen*, Bulletin de l'Académie des Sciences de l'URSS, Classe des sciences mathématiques et naturelles, series 7, 1932, pp. 1163-1202; (b) *Über einige Funktionen, die in gegebenen Intervallen am wenigsten von Null abweichen*, Bulletin de la Société Physico-Mathématique de Kazan, series 3, vol. 3 (1928), pp. 1-69.

⁶ Zolotareff and Achyesser make extensive use of the theory of elliptic functions; however, W. Markoff does not.

2. On the polynomials of Zolotareff

1. It is a classical fact that there is a unique polynomial $T_n(x)$ of degree n (the n^{th} polynomial of Tchebycheff) having the following property: The curve $y = T_n(x)$, $-1 \leq x \leq +1$, consists of n monotonic arcs varying between $+1$ and -1 ; $T_n(1) = 1$, and $T_n(-1) = (-1)^n$. This polynomial satisfies the differential equation

$$(2.1) \quad n^2(1 - y^2) = (1 - x^2)y'^2$$

from which follows

$$(2.2) \quad y = \cos \left\{ n \int_1^x (1 - t^2)^{-1/2} dt \right\}.$$

2. We show that there are infinitely many polynomials y of degree n possessing the following property: The curve y , $-1 \leq x \leq +1$, consists of $n - 1$ monotonic arcs varying between $+1$ and -1 , $y = 1$ for $x = 1$, and $y = (-1)^{n-1}$ for $x = -1$. Such a curve necessarily has $n - 1$ roots in $-1 \leq x \leq +1$ and consequently one more outside this interval. If this additional root is > 1 , y satisfies a differential equation of the form

$$(2.3) \quad n^2(1 - y^2) = (1 - x^2)y'^2 \frac{(B - x)(C - x)}{(A - x)^2}$$

where $y' = 0$ for $x = A$, $y = 1$ for $x = B$, $y = -1$ for $x = C$, and $1 < A < B < C$. A similar differential equation holds if the additional root of y mentioned above is < -1 . (The second case can be obtained from the first one by replacing x by $-x$.)

Solving the differential equation (2.3), we obtain

$$(2.4) \quad y = \cos \left\{ n \int_1^x (A - t)(B - t)^{-1/2}(C - t)^{-1/2}(1 - t^2)^{-1/2} dt \right\}.$$

From the properties of y mentioned above we find

$$(2.5) \quad \int_{-1}^{+1} (A - t)(B - t)^{-1/2}(C - t)^{-1/2}(1 - t^2)^{-1/2} dt = (n - 1)\pi/n,$$

$$(2.6) \quad \int_{+1}^B (A - t)(B - t)^{-1/2}(C - t)^{-1/2}(t^2 - 1)^{-1/2} dt = 0,$$

$$(2.7) \quad \int_B^C (t - A)(t - B)^{-1/2}(C - t)^{-1/2}(t^2 - 1)^{-1/2} dt = \pi/n.$$

By a well-known application of Cauchy's theorem we see that the sum of the integrals (2.5) and (2.7) is π , so that (2.7) is a consequence of (2.5).

Conversely, if (2.5) and (2.6) hold, an easy discussion (encircling the singular points -1 , $+1$, B , C) shows that (2.4) is an analytic function single-valued and regular in the whole finite x -plane. If $x \rightarrow \infty$ we find $y = O(|x|^{-n})$, so that y

must be a polynomial of degree n . Of course it satisfies the differential equation (2.3), and it has all the properties mentioned above.

For later purposes we note that

$$(2.8) \quad y' = n^2 \frac{(A-1)^2}{(B-1)(C-1)} \quad \text{at } x = 1,$$

$$(2.9) \quad (-1)^n y' = n^2 \frac{(A+1)^2}{(B+1)(C+1)} \quad \text{at } x = -1.$$

These values can be obtained from the differential equation (2.3).

3. LEMMA 1. *Of the three quantities A, B, C ($1 < A < B < C$) satisfying the two transcendental equations (2.5) and (2.6), any one can be prescribed arbitrarily provided that*

$$(2.10) \quad A > 1, \text{ or } B > 1, \text{ or } C > c_n = 1 + 2\alpha_n = 1 + 2 \tan^2 [\pi/(2n)]$$

respectively; the two others are then uniquely determined. As A increases monotonically from 1 to $+\infty$, B and C increase likewise from 1 to $+\infty$ and from c_n to $+\infty$, respectively.

Furthermore the values of $y, y', \dots, y^{(n)}$ for a fixed x not less than one, and the values of $(-1)^n y, (-1)^{n-1} y', \dots, y^{(n)}$ for a fixed x not greater than -1 , are increasing functions of A .

The only exceptions are $y = 1$ for $x = 1$ and $(-1)^n y = -1$ for $x = -1$. In particular, the expressions (2.8) and (2.9) are respectively increasing and decreasing functions of A .

In order to prove this Lemma, let B denote a fixed value, greater than 1, and let C be variable, such that $C > B$; we define $A = A(C)$ by (2.6) so that $1 < A < B$. Then

$$\int_1^B \left(\frac{dA}{dC} - \frac{1}{2} \frac{A-t}{C-t} \right) (B-t)^{-1} (C-t)^{-1} (t^2-1)^{-1} dt = 0;$$

hence

$$\frac{dA}{dC} = \frac{1}{2} \frac{A-t_0}{C-t_0}, \quad \text{where } 1 < t_0 < B.$$

Now consider the function $\lambda(C)$ defined by the left-hand member of (2.5), where $A = A(C)$. We find

$$\begin{aligned} \lambda'(C) &= \int_{-1}^{+1} \left(\frac{dA}{dC} - \frac{1}{2} \frac{A-t}{C-t} \right) (B-t)^{-1} (C-t)^{-1} (1-t^2)^{-1} dt \\ &= \int_{-1}^{+1} \left(\frac{1}{2} \frac{A-t_0}{C-t_0} - \frac{1}{2} \frac{A-t}{C-t} \right) (B-t)^{-1} (C-t)^{-1} (1-t^2)^{-1} dt < 0, \end{aligned}$$

so that $\lambda(C)$ is decreasing. Let $C \rightarrow B$; then from (2.6) we see that $A \rightarrow B$, so that

$$\lambda(C) \rightarrow \int_{-1}^{+1} (1 - t^2)^{-1} dt = \pi.$$

Since $\lim_{C \rightarrow \infty} \lambda(C) = 0$, the equation $\lambda(C) = (n - 1)\pi/n$ has precisely one solution.⁷

4. Further let $p(x)$ and $q(x)$ be two special cases of (2.4) corresponding to the values A', B', C' and A'', B'', C'' of A, B, C , respectively. First suppose that $p'(1) < q'(1)$. Considering the polynomial $\delta(x) = p(x) - q(x)$ at the n points in $-1 \leq x \leq +1$ at which $p(x) = \pm 1$ and assuming that $\delta(x) \neq 0$, a familiar argument furnishes the existence of $n - 1$ distinct points $+1 > \eta_1 > \eta_2 > \dots > \eta_{n-1} > -1$ such that $\delta'(\eta_1) > 0, \delta'(\eta_2) < 0, \dots$. Furthermore $\delta'(1) < 0$, so that $\delta'(x)$ has $n - 1$ roots (that is, all its roots) in $-1 < x < +1$. The same holds for $\delta''(x), \delta'''(x), \dots$ so that $\delta(x) < 0, \delta'(x) < 0, \delta''(x) < 0, \dots$ for $x \geq 1$ [except that $\delta(1) = 0$], and also $(-1)^n \delta(x) < 0, (-1)^{n-1} \delta'(x) < 0, (-1)^{n-2} \delta''(x) < 0, \dots$ for $x \leq -1$ [except that $\delta(-1) = 0$]. From this we easily conclude that the relations $A' < A'', B' < B'', C' < C''$ hold for the constants corresponding to $p(x)$ and $q(x)$.

If $p'(1) = q'(1)$ the previous argument still holds good [unless $\delta(x) \equiv 0$], except that $\delta'(1) = 0$ so that the roots of $\delta'(x)$ are in $-1 < x \leq +1$. Consequently $\delta'(x) < 0$ for $x > 1$. Interchanging $p(x)$ and $q(x)$ we obtain $\delta'(x) > 0, x > 1$, which is a contradiction; so that in this case $p(x) \equiv q(x), A' = A'', B' = B'', C' = C''$.

From the previous considerations we conclude that B and C are increasing functions of A . It remains to calculate the limits of B and C as $A \rightarrow 1$ and $A \rightarrow +\infty$. In the former case, (2.6) shows that $B \rightarrow 1$, and from (2.5) we obtain $C \rightarrow c_n$ since the equation

$$\int_{-1}^{+1} (1 + t)^{-1} (\gamma - t)^{-1} dt = (n - 1)\pi/n$$

has the unique solution $\gamma = c_n$. If $A \rightarrow +\infty$ it is obvious that $B \rightarrow +\infty, C \rightarrow +\infty$. This completes the proof of Lemma 1.

5. In what follows we denote the polynomial (2.4) [for which (2.5) and (2.6) hold] by $y = u_n(x; A)$. We note that, from (2.4) and (2.10),

$$\begin{aligned} u_n(x; +1) &= \lim_{A \rightarrow 1} u_n(x; A) \\ (2.11) \quad &= \cos \left\{ n \int_1^x (1 + t)^{-1} (c_n - t)^{-1} dt \right\} = -T_n \left(\frac{x - c_n}{1 + c_n} \right). \end{aligned}$$

Hence $u'_n(+1; +1) = 0$. Also

$$(2.12) \quad u''_n(+1; +1) = -(1 + \alpha_n)^{-2} T''_n[\cos(\pi/n)] = -\frac{1}{4} n^2 \cot^2[\pi/(2n)].$$

⁷ These considerations require only slight modifications if we replace the right-hand members in (2.5) and (2.6) by $\nu\pi/n$ and $\pi - \nu\pi/n, 1 \leq \nu \leq n - 1$. The resulting polynomials have been used for various purposes by Achyzer; see loc. cit.

Further, let $A \rightarrow +\infty$ so that $B \rightarrow +\infty$ and $C \rightarrow +\infty$. From (2.5)

$$(A - t_1)(B - t_1)^{-1}(C - t_1)^{-1} = (n - 1)/n$$

where t_1 is a suitably chosen number between 0 and 1. Hence $A(BC)^{-1} \rightarrow 1 - 1/n$, so that, from (2.4),

$$(2.13) \quad u_n(x; +\infty) = \lim_{A \rightarrow \infty} u_n(x; A) = T_{n-1}(x).$$

Hence

$$(2.14) \quad u'_n(+1; +\infty) = (n - 1)^2;$$

$$(2.15) \quad u''_n(+1; +\infty) = \frac{1}{3}n(n - 1)^2(n - 2).$$

Therefore, as A increases from 1 to $+\infty$, $u'_n(+1; A)$ increases from 0 to $(n - 1)^2$, and $u''_n(+1; A)$ increases from the negative value (2.12) to the positive value (2.15), corresponding respectively to $A = 1$ and $A = +\infty$. There is precisely one value of A , $A = A_n$, for which $u''_n(+1; A_n) = 0$. We denote the corresponding values of B and C by B_n and C_n . In §§4 and 5 we shall prove that the function $u_n(x; A_n)$ furnishes the solution of I. Schur's problem formulated above, provided n is sufficiently large.

From the differential equation (2.3) we obtain

$$(2.16) \quad u''_n(+1; A) = \frac{1}{3}n^2 \frac{(A - 1)^2}{(B - 1)(C - 1)} \left\{ n^2 \frac{(A - 1)^2}{(B - 1)(C - 1)} - 1 - 2 \left(\frac{2}{A - 1} - \frac{1}{B - 1} - \frac{1}{C - 1} \right) \right\},$$

so that the condition $u''_n(+1; A) = 0$ is equivalent to

$$(2.17) \quad n^2 \frac{(A - 1)^2}{(B - 1)(C - 1)} = 1 + 2 \left(\frac{2}{A - 1} - \frac{1}{B - 1} - \frac{1}{C - 1} \right).$$

The transcendental equations (2.5), (2.6) and (2.17) determine the constants $A = A_n$, $B = B_n$, $C = C_n$ uniquely. These constants depend only on n .

The polynomial $y = u_n(x; A_n)$ is completely determined by the following conditions: The curve y , $-1 \leq x \leq +1$, consists of $n - 1$ monotonic arcs varying between $+1$ and -1 , $y = 1$ for $x = 1$, $y = (-1)^{n-1}$ for $x = -1$ and $y'' = 0$ for $x = 1$.

3. The limiting process $n \rightarrow \infty$

1. First we prove the following

LEMMA 2. The constants A_n , B_n , C_n defined by the transcendental equations (2.5), (2.6), (2.17) satisfy

$$(3.1) \quad \begin{aligned} \lim_{n \rightarrow \infty} n^2(A_n - 1) &= a^2/2, & \lim_{n \rightarrow \infty} n^2(B_n - 1) &= b^2/2, \\ \lim_{n \rightarrow \infty} n^2(C_n - 1) &= c^2/2 \end{aligned}$$

where $0 < a < b < c$. The numerical values of a , b , c are given in (3.17).

By Lemma 1

$$\liminf_{n \rightarrow \infty} n^2(C_n - 1) \geq \pi^2/2$$

and from (2.17)

$$(3.2) \quad n^2 \frac{A_n - 1}{C_n - 1} \geq \frac{2}{A_n - 1} - \frac{2}{C_n - 1},$$

$$n^4(A_n - 1)^2 + 2n^2(A_n - 1) \geq 2n^2(C_n - 1),$$

so that

$$(3.3) \quad \liminf_{n \rightarrow \infty} n^2(A_n - 1) \geq (1 + \pi^2)^{\frac{1}{2}} - 1.$$

The same inequality holds if we replace A_n by B_n .

On the other hand, let us assume that $n^2(C_n - 1) \rightarrow +\infty$ for a proper subsequence $n = n_\nu$ as $\nu \rightarrow \infty$; then, from (3.2), $n^2(A_n - 1) \rightarrow +\infty$, so that $n^2(B_n - 1) \rightarrow +\infty$. Therefore, by (2.17), for the same subsequence $n = n_\nu$,

$$(3.4) \quad \frac{(A_n - 1)^2}{(B_n - 1)(C_n - 1)} \rightarrow 0,$$

Now let ω be a fixed positive number; for large n , from (2.5),

$$(3.5) \quad \begin{aligned} \pi &= n \int_{-1}^{+1} (1 - t^2)^{-\frac{1}{2}} dt \{1 - (A_n - t)(B_n - t)^{-1}(C_n - t)^{-1}\} \\ &> n \int_0^{\omega/n} d\varphi \{1 - (A_n - \cos \varphi)(B_n - \cos \varphi)^{-1}(C_n - \cos \varphi)^{-1}\} \\ &= \int_0^\omega d\psi \{1 - (A_n - \cos(\psi/n))(B_n - \cos(\psi/n))^{-1}(C_n - \cos(\psi/n))^{-1}\}. \end{aligned}$$

Since

$$\begin{aligned} &(A_n - \cos(\psi/n))(B_n - \cos(\psi/n))^{-1}(C_n - \cos(\psi/n))^{-1} \\ &\leq (A_n - 1)(B_n - 1)^{-1}(C_n - 1)^{-1} \left\{ 1 + \frac{1 - \cos(\psi/n)}{A_n - 1} \right\} \end{aligned}$$

and since for $n = n_\nu$, as $\nu \rightarrow \infty$,

$$\frac{n^2(1 - \cos(\psi/n))}{n^2(A_n - 1)} \rightarrow 0$$

uniformly in ψ , for $0 \leq \psi \leq \omega$, we find $\pi \geq \omega$. This is a contradiction if we choose $\omega > \pi$. Thus we have proved that the points of accumulation of the sequences $n^2(A_n - 1)$, $n^2(B_n - 1)$, $n^2(C_n - 1)$ are positive and finite.

2. Now let $n = n_\nu$ be a subsequence for which the limits (3.1) exist, where

$0 < a \leq b \leq c < +\infty$. From (2.5), (2.6) and (2.7) we shall derive $a < b < c$ and

$$(3.6) \quad \int_0^\infty \{1 - (a^2 + u^2)(b^2 + u^2)^{-1}(c^2 + u^2)^{-1}\} du = \pi,$$

$$(3.7) \quad \int_0^b (a^2 - u^2)(b^2 - u^2)^{-1}(c^2 - u^2)^{-1} du = 0,$$

$$(3.8) \quad \int_b^c (u^2 - a^2)(u^2 - b^2)^{-1}(c^2 - u^2)^{-1} du = \pi.$$

Also from (2.17) by the same limiting process ($n = n_\nu$, $\nu \rightarrow \infty$),

$$(3.9) \quad \frac{a^4}{b^2 c^2} - 4 \left(\frac{2}{a^2} - \frac{1}{b^2} - \frac{1}{c^2} \right) = 0.$$

Instead of (3.7) we can show more precisely

$$(3.10) \quad \left\{ \begin{array}{l} n \int_1^{A_n} (A_n - t)(B_n - t)^{-1}(C_n - t)^{-1}(t^2 - 1)^{-1} dt \\ \qquad \qquad \qquad \rightarrow \int_0^a (a^2 - u^2)(b^2 - u^2)^{-1}(c^2 - u^2)^{-1} du, \\ n \int_{A_n}^{B_n} (t - A_n)(B_n - t)^{-1}(C_n - t)^{-1}(t^2 - 1)^{-1} dt \\ \qquad \qquad \qquad \rightarrow \int_a^b (u^2 - a^2)(b^2 - u^2)^{-1}(c^2 - u^2)^{-1} du, \end{array} \right.$$

the two limits being the same.

First, (3.9) is obvious and this equation shows that $a = b = c$ is impossible. In case $a < b = c$ both formulas (3.10) follow easily [writing $t = 1 + u^2/(2n^2)$]; but the first limit is finite and the second one turns out to be $+\infty$, which is a contradiction. In case $a = b < c$ the same formulas can be easily established again, but the first limit is positive whereas the second one is 0 [since $\max\{(t - A_n)(C_n - t)^{-1}(t^2 - 1)^{-1}\}$, $A_n \leq t \leq B_n$, is bounded]. Therefore $a < b < c$.

Now (3.7) and (3.8) follow directly, and (3.6) can also be easily obtained. However (3.6) follows also from (3.8) by applying Cauchy's theorem to

$$f(z) = 1 - (a^2 - z^2)(b^2 - z^2)^{-1}(c^2 - z^2)^{-1}$$

integrated along the half-circle $|z| = R$, $\Re z \geq 0$ and along the segment $\Re z = 0$, $-R \leq \Im z \leq +R$, $R \rightarrow +\infty$.

3. Substituting $u^2 = b^2 \sin^2 \varphi$ in (3.7) and $u^2 = c^2 - (c^2 - b^2) \sin^2 \varphi$ in (3.8) we find

$$(3.11) \quad \int_0^{\pi/2} (a^2 - b^2 \sin^2 \varphi)(c^2 - b^2 \sin^2 \varphi)^{-1} d\varphi = 0,$$

$$(3.12) \quad \int_0^{\pi/2} \{c^2 - a^2 - (c^2 - b^2) \sin^2 \varphi\} \{c^2 - (c^2 - b^2) \sin^2 \varphi\}^{-1} d\varphi = \pi.$$

Using the standard notation these equations can be written in the form

$$(3.13) \quad (1 - a^2/c^2)K = E, \quad cE' - (a^2/c)K' = \pi$$

where the complete elliptic integrals K and E belong to the modulus $k = b/c$. Eliminating a^2/c^2 we find

$$(3.14) \quad E/K + (E' - \pi/c)/K' = 1.$$

Comparing this with the classical equation⁸

$$(3.15) \quad EK' + E'K - KK' = \pi/2$$

we obtain $c = 2K$. Hence

$$(3.16) \quad a^2 = 4K(K - E), \quad b = 2kK, \quad c = 2K.$$

The relation (3.9) furnishes the transcendental equation (1.5) of Theorem 1 (see §1) for the modulus k . This equation has precisely one root as k^2 goes from 0 to 1 [which shows that the limits (3.1) exist as $n \rightarrow \infty$ unrestrictedly]. Indeed, differentiating the left-hand member of (1.5) with respect to k^2 ,⁹ we have

$$\frac{3}{2}E\{k'^{-2}(K - E)^2 - 1\},$$

where k' is the complementary modulus. The expression in the curly bracket increases with k^2 , as the well-known power series expansion of K and E shows; it is negative for small k^2 and positive as k^2 approaches 1. Therefore the left-hand member of (1.5) first decreases and then increases; but for $k^2 = 0$ it is zero and for $k^2 \rightarrow 1 - 0$ it tends to $+\infty$. This establishes Lemma 2.

Using the tables of Milne-Thomson¹⁰ we find

$$(3.17) \quad \begin{aligned} k^2 &= 0.84 \dots, & a^2 &= 11.4055 \dots, & b &= 4.3245 \dots, \\ & & c &= 4.7185 \dots, & a^4/b^2c^2 &= 0.3124 \dots. \end{aligned}$$

We also note that (2.4) implies that

$$(3.18) \quad \lim_{n \rightarrow \infty} u_n(\cos(z/n); A_n) = \cos \left\{ \int_0^z (a^2 + u^2)(b^2 + u^2)^{-1}(c^2 + u^2)^{-1} du \right\}$$

uniformly in z , for all complex z such that $|z| \leq R$.

4. Another limiting formula important for the proof of Theorem 1, is

LEMMA 3. Let $A = A'_n$ be a sequence of values such that $A'_n - 1 = o(n^{-2})$.

⁸ See for instance, E. T. Whittaker and G. N. Watson, *A course of Modern Analysis*, Fourth edition, 1935, p. 520.

⁹ See Whittaker-Watson, loc. cit. p. 521.

¹⁰ L. M. Milne-Thomson, *Ten-figure table of the complete elliptic integrals K, K', E, E' and a table of $1/\partial^2_1(0|\tau), 1/\partial^2_2(0|\tau')$* , Proceedings of the London Mathematical Society, series 2, vol. 33 (1932), pp. 160-164.

Denoting the corresponding values of B and C determined from the equations (2.5) and (2.6) by B'_n and C'_n , respectively, we have

$$(3.19) \quad \lim_{n \rightarrow \infty} u_n(\cos(z/n); A'_n) = -\cos\{(\pi^2 + z^2)^{\frac{1}{2}}\}.$$

The last equation holds uniformly in z , for all complex z such that $|z| \leq R$.

We note that (3.19) arises from (3.18) on writing $a = b = 0$, $c = \pi$.

For the proof we use an argument similar to that of Part 1. Let ω be fixed, $\omega > 0$; we find [see (3.5)]

$$(3.20) \quad \pi > \int_0^\omega d\psi \{1 - (A'_n - \cos(\psi/n))(B'_n - \cos(\psi/n))^{-1}(C'_n - \cos(\psi/n))^{-1}\}.$$

Assuming for a certain subsequence $n = n_\nu$, $\nu \rightarrow \infty$, that the limits

$$\lim n^2(B'_n - 1) = \beta, \quad \lim n^2(C'_n - 1) = \gamma$$

exist, we have $\beta \geq 0$, $\gamma \geq \pi^2/2$. Thus we conclude from (3.20)

$$\pi \geq \int_0^\omega d\psi \{1 - (\psi^2/2)(\beta + \psi^2/2)^{-1}(\gamma + \psi^2/2)^{-1}\},$$

so that

$$(3.21) \quad \pi \geq \int_0^\infty d\psi \{1 - (\psi^2/2)(\beta + \psi^2/2)^{-1}(\gamma + \psi^2/2)^{-1}\}.$$

Now

$$(3.22) \quad \pi = \int_0^\infty d\psi \{1 - \psi(\pi^2 + \psi^2)^{-1}\};$$

consequently (3.21) and (3.22) involve a contradiction, unless $\beta = 0$, $\gamma = \pi^2/2$.

Further

$$(3.23) \quad \begin{aligned} & u_n(\cos(z/n); A'_n) \\ &= \cos \left\{ \int_0^z (A'_n - \cos(\psi/n))(B'_n - \cos(\psi/n))^{-1}(C'_n - \cos(\psi/n))^{-1} d\psi \right\}. \end{aligned}$$

Now let $0 < \epsilon < \pi < R$ and $|z| = R$. Then

$$\begin{aligned} & \int_0^\epsilon (A'_n - \cos(\psi/n))(B'_n - \cos(\psi/n))^{-1}(C'_n - \cos(\psi/n))^{-1} d\psi \\ & \leq \int_0^\epsilon (A'_n - \cos(\psi/n))^{-1}(C'_n - \cos(\psi/n))^{-1} d\psi \rightarrow \int_0^\epsilon \psi(\pi^2 + \psi^2)^{-1} d\psi \end{aligned}$$

as $n \rightarrow \infty$; the last integral is arbitrarily small with ϵ . Integrating from ϵ to z , we can assume that $\psi \neq 0, \pm \pi$ on the path of integration; and the assertion follows immediately from (3.23) for $n \rightarrow \infty$.

4. Proof of Theorem 1

In what follows, the symbols Q_n , $Q_n(x_0)$ defined in §1 are used.

1. LEMMA 4. Suppose $-1 \leq x_0 \leq +1$, and let $f_0(x)$ be a polynomial of the class $Q_n(x_0)$ for which $\max |f'_0(x)|$, $f(x) \in Q_n(x_0)$, is attained. Then $|f_0(x)|$ assumes its maximum 1 at least n times in $-1 \leq x \leq +1$.

The proof follows the usual lines. Let $f'_0(x_0) > 0$ and let us suppose that the assertion of Lemma 4 does not hold. Denote by x_1, x_2, \dots, x_l ; $l < n$, the distinct values in $-1 \leq x \leq +1$ for which $|f_0(x_r)| = 1$ and write $\omega(x) = \prod_{r=1}^l (x - x_r)$. If $-1 < x_0 < +1$ we have $x_0 \neq x_r$ [otherwise $f'_0(x_0)$ would be 0]. However if $x_0 = \pm 1$ we may have $x_0 = x_r$, in which case $\omega(x_0) = 0$ but $\omega'(x_0) \neq 0$.

We form the polynomial

$$(4.1) \quad r(x) = -\sum_{r=1}^l \operatorname{sgn} f_0(x_r) \frac{\omega(x)}{\omega'(x_r)(x - x_r)} + \omega(x)\{a(x - x_0) + b\}$$

and want to determine the constants a and b such that $r'(x_0) > 0$, $r''(x_0) = 0$; this can certainly be done provided the linear equations

$$a\omega(x_0) + b\omega'(x_0) = G,$$

$$2a\omega'(x_0) + b\omega''(x_0) = H$$

have a determinant $\neq 0$. Now $\omega(x_0)\omega''(x_0) - 2\{\omega'(x_0)\}^2 \neq 0$ is obvious if $\omega(x_0) = 0$ (cf. above); but if $\omega(x_0) \neq 0$,

$$\frac{\omega''(x_0)}{\omega(x_0)} - 2\left\{\frac{\omega'(x_0)}{\omega(x_0)}\right\}^2 = -\sum_{r=1}^l (x_0 - x_r)^{-2} - \left\{\frac{\omega'(x_0)}{\omega(x_0)}\right\}^2 < 0.$$

Obviously $r(x)$ is of degree $l+1 \leq n$ and we find for sufficiently small $\epsilon > 0$ that $|f_0(x) + \epsilon r(x)| \leq 1$ in $-1 \leq x \leq +1$; hence $f_0(x) + \epsilon r(x)$ belongs to $Q_n(x_0)$. On the other hand $f'_0(x_0) + \epsilon r'(x_0) > f'_0(x_0)$ which is a contradiction. This proves Lemma 4.

2. Let the extremum (1.2) be attained for the value x_0 and for $f(x) = f_0(x)$, $f_0(x) \in Q_n(x_0)$. Then $f''_0(x_0) = 0$, and $f_0(x)$ possesses the property formulated in Lemma 4. Further we show that $f'''_0(x_0) \neq 0$. By Lemma 4, $|f_0(x)|$ attains its relative maximum 1 in $-1 < x < +1$ for at least $n - 2$ distinct points for which $f'_0(x) = 0$. Since $f''_0(x)$ vanishes an odd number of times between two consecutive roots of $f'_0(x)$, we find that $f''_0(x)$ has precisely one simple root between two consecutive roots of $f'_0(x)$, and these roots of $f''_0(x)$ are maximum points of $|f'_0(x)|$. The number of these maximum points is at least $n - 3$. If x_0 is one of these points, we must have $f'''_0(x_0) \neq 0$. If x_0 is different from these maximum points (whose number in this case is $n - 3$), then we must have again $f'''_0(x_0) \neq 0$, and thus there is a relative maximum of $|f'_0(x)|$ at $x = x_0$.

If we assume that $f'_0(x_0) > 0$ then $f''_0(x_0) = 0$, $f'''_0(x_0) < 0$, so that $f'_0(x)$ has a relative maximum at $x = x_0$.

Now we distinguish various cases.

(a) $x_0 = \pm 1$.

Let $x_0 = +1$ and let us denote an extremum polynomial of our problem by $u_n(x)$, $u'_n(1) > 0$, $u''_n(1) = 0$. As we showed before, $u'_n(x)$ has at least $n - 2$ and $u''_n(x)$ at least $n - 3$ distinct roots in $-1 < x < +1$. Since $u''_n(1) = 0$, we find that $n - 2$ is the precise number of roots of $u'_n(x)$ in $-1 < x < +1$. Consequently $|u_n(-1)| = |u_n(+1)| = 1$; and, since $u'_n(1) > 0$, we find $u_n(1) = 1$, $u_n(-1) = (-1)^{n-1}$.

Thus the curve $y = u_n(x)$, $-1 \leq x \leq +1$, consists of $n - 1$ monotonic arcs varying between $+1$ and -1 , and $u_n(1) = 1$, $u_n(-1) = (-1)^{n-1}$, $u'_n(1) > 0$, $u''_n(1) = 0$.

Hence from the last remark of §2 we conclude that $u_n(x)$ is identical with the polynomial $u_n(x; A_n)$ defined there.

Consequently, under the assumption $x_0 = \pm 1$, the extremum polynomials of our problem are $\pm u_n(x; A_n)$ and $\pm u_n(-x; A_n)$, respectively. The asymptotic value of $|u'_n(1; A_n)|$ is $a^4 b^{-2} c^{-2} \cdot n^2$ [see (3.17)].

(b) $-1 < x_0 < +1$, and there exists a polynomial $g(x)$ of Q_n for which $|g'(x_0)| > |f'_0(x_0)|$. Suppose $f'_0(x_0) > 0$, $g'(x_0) > 0$.

Consider the polynomial $h_\epsilon(x) = f_0(x) + \epsilon\{g(x) - f_0(x)\}$, $0 < \epsilon < 1$. Obviously $h_\epsilon(x) \in Q_n$; furthermore $h'_\epsilon(x_0) > f'_0(x_0)$. For sufficiently small ϵ there is a root of $h'_\epsilon(x)$ in the neighborhood of x_0 , x'_0 say, and $h'_\epsilon(x)$ attains a positive relative maximum at $x = x'_0$. We evidently have

$$h'_\epsilon(x'_0) \geq h'_\epsilon(x_0) > f'_0(x_0)$$

which shows that $f_0(x)$ cannot be the extremum polynomial.

5. Proof of Theorem 1 (continued)

The remaining case requires a more elaborate discussion. This case is:

(c) $-1 < x_0 < +1$ and $f_0(x)$ is the polynomial in Q_n with the maximum value of $f'_0(x_0)$.

Then W. Markoff has shown¹¹ that $f_0(x)$ must be one of the polynomials

$$(5.1) \quad \begin{aligned} &\pm T_n(x), \quad \pm T_{n-1}(x), \quad \pm T_n\left(\frac{x - \alpha}{1 + \alpha}\right), \quad \pm T_n\left(\frac{x + \alpha}{1 + \alpha}\right), \\ &\pm u_n(\pm x; A) \end{aligned}$$

where $0 < \alpha < \alpha_n = \tan^2[\pi/(2n)]$, and $u_n(x; A)$ are the Zolotareff polynomials defined and discussed above. As $n \rightarrow \infty$, the largest relative maximum of $|T_n(x)|$ in $-1 < x < +1$ is asymptotically $M \cdot n^2$ where $-M$ is the minimum of $\sin \theta / \theta$ for real θ , that is $M = 0.2172 \dots$. Comparing this result with the asymptotic value of $u'_n(1; A_n)$, that is with $a^4 b^{-2} c^{-2} \cdot n^2$ [see (3.17)], we see that for large values of n the four first types in (5.1) can be excluded.

¹¹ Loc. cit. p. 249.

As W. Markoff has further shown¹², $f_0(x) = \pm u_n(x; A)$ if and only if (a) x_0 belongs to certain open intervals in $-1 < x < +1$, and (b):

$$(5.2) \quad \frac{d}{dx} \left\{ \frac{(1-x^2)u'_n(x; A)}{x-A} \right\} = 0 \quad \text{at } x = x_0.$$

Since $u'_n(x_0; A) \neq 0$, and $u''_n(x_0; A) = 0$, the latter-mentioned condition implies that

$$(5.3) \quad x_0 = A - (A^2 - 1)^{\frac{1}{2}}, \quad A - 1 = (1 - x_0)^2 / (2x_0),$$

so that $0 < x_0 < +1$. Now we distinguish again two cases:

(c') $0 < x_0 \leq (1 - 16n^{-2})^{\frac{1}{2}}$. According to S. Bernstein's theorem

$$(5.4) \quad |u'_n(x; A)| \leq n(1 - x^2)^{-\frac{1}{2}} \leq n^{\frac{1}{2}}.$$

(c'') $(1 - 16n^{-2})^{-\frac{1}{2}} < x_0 < 1$. Then $A - 1 = A'_n - 1 = O(n^{-4})$. Now we assume that this case occurs for an infinite number of values of n , and we write $x_0 = \cos(z_0/n)$; then z_0 is bounded. From Lemma 3 we conclude that

$$(5.5) \quad \lim n^{-2} u'_n(\cos(z/n); A'_n) = -\frac{\sin\{(\pi^2 + z^2)^{\frac{1}{2}}\}}{(\pi^2 + z^2)^{\frac{1}{2}}}.$$

The maximum of the absolute value of the last expression for real z is $M = 0.2172 \dots$ so that this case can be also eliminated.

The assumption $f_0(x) = \pm u_n(-x; A)$ can be dealt with similarly.

Thus for large n only Case (a) remains. This completes the proof of Theorem 1.

6. Proof of Theorem 2

1. First we consider again the case (c) defined in §5 and let x_0 belong to one of the open intervals in $-1 \leq x \leq +1$ in which the maximum of $f'(x_0), f(x) \in Q_n$, is attained for the Zolotareff polynomial $f(x) = u_n(x; A)$. [The argument is similar for $-u_n(x; A)$ or $\pm u_n(-x; A)$.] Then $f(x) = u_n(x; A) = f_0(x)$, where $f_0(x)$ has the same meaning as in §§4 and 5, so that $f_0(x) \in Q_n(x_0)$; that is, $f''_0(x_0) = 0$. We have $f'_0(x_0) > 0$, $f'''_0(x_0) < 0$.

By an important theorem of W. Markoff¹³, to every positive ϵ correspond values x_1 such that¹⁴

(α) $0 < |x_1 - x_0| < \epsilon$;

(β) if $f_1(x) = u_n(x; A')$ denotes the polynomial of Q_n for which $f'(x_1)$ becomes a maximum, then

$$(6.1) \quad f'_1(x_1) > f'_0(x_0).$$

¹² Loc. cit. pp. 233-246.

¹³ Loc. cit. p. 257.

¹⁴ In fact, a whole half-neighborhood of x_0 satisfies this condition.

Now if ϵ is sufficiently small, $f_1''(x)$ will have a root, say x_1' , in the neighborhood of x_0 ; we can assume that $-1 < x_1' < +1$. Also $f_1'''(x_1') < 0$, so that $f_1'(x)$ has a relative maximum at $x = x_1'$; hence

$$(6.2) \quad f_1'(x_1') \geq f_1'(x_1) > f_0'(x_0),$$

which shows that $f_0(x)$ can not be the solution of our problem.

This argument leaves as the only possibilities for $f_0(x)$ either the Zolotareff polynomials $\pm u_n(x; A_n)$ with $x_0 = \pm 1$, or the Tchebycheff polynomials $\pm T_n(x)$.

2. Let D_n be the largest root of $u_n(x; A_n)$, $B_n < D_n < C_n$. Using the convexity of $u_n(x; A_n)$ for $x > 1$, we deduce

$$(6.3) \quad D_n - B_n > C_n - D_n.$$

Further we make use of a theorem of I. Schur on the largest roots of the derivatives of an algebraic equation with only real roots.¹⁵ Applying this theorem to $u_n(x; A_n)$ we obtain

$$(6.4) \quad D_n - A_n \leq A_n - 1$$

so that

$$2(A_n - 1) \geq D_n - 1 > \frac{1}{2}(B_n - 1 + C_n - 1) > \{(B_n - 1)(C_n - 1)\}^{\frac{1}{2}}.$$

Hence, from (2.8),

$$(6.5) \quad u_n'(1; A_n) > n^2/4.$$

3. On the other hand we show that

$$(6.6) \quad |T_n'(x)| \leq n^2/4 \quad \text{if} \quad T_n''(x) = 0$$

provided $n \geq 5$ (with equality only if $n = 5$). Incidentally, I. Schur has proved (6.6) for all large n .¹⁶

Let φ be a root of the equation $\tan n\varphi = n \tan \varphi$, $0 < \varphi < \pi/2$. Then the assertion is

$$(6.7) \quad n \left| \frac{\sin n\varphi}{\sin \varphi} \right| = n^2(n^2 \sin^2 \varphi + \cos^2 \varphi)^{-\frac{1}{2}} \leq n^2/4, \quad \sin \varphi \geq \left(\frac{15}{n^2 - 1} \right)^{\frac{1}{2}}.$$

It is sufficient to show this for the largest root $x_n = \cos \varphi_n$ of $T_n''(x)$; that is, for the smallest positive value φ_n , $\pi < n\varphi_n < 3\pi/2$, satisfying the equation above.

The function

$$(6.8) \quad h(\psi) = \frac{\tan n\psi}{n \tan \psi}$$

¹⁵ I. Schur, *Zwei Sätze über algebraische Gleichungen mit lauter reellen Wurzeln*, Journal für die reine und angewandte Mathematik, vol. 144 (1914), pp. 75-88.

¹⁶ I. Schur, loc. cit.², p. 277.

increases from 0 to $+\infty$ as ψ increases from π/n to $3\pi/(2n)$. Let z be the smallest positive root of the equation $\tan z = z$, $\pi < z < 3\pi/2$. Since

$$(6.9) \quad h(z/n) = \frac{z}{n \tan(z/n)} < 1,$$

we have $\varphi_n > z/n$, so that (6.7) follows from

$$(6.10) \quad \sin(z/n) \geq \left(\frac{15}{n^2 - 1} \right)^{\frac{1}{2}}.$$

Since $n \sin(z/n)$ increases and $n^2/(n^2 - 1)$ decreases as n increases, the last inequality will be proved for $n \geq 6$ if we prove it for $n = 6$. But

$$(6.11) \quad \sin(z/6) \geq (3/7)^{\frac{1}{2}} = 0.6546 \dots,$$

since¹⁷ $z = 4.4934 \dots$ and $\sin(z/6) = 0.6808 \dots$.

In the case $n = 5$ we have

$$(6.12) \quad T_5'(x) = 320x^3 - 120x, \quad x_5 = \cos \varphi_5 = (3/8)^{\frac{1}{2}}, \quad \sin \varphi_5 = (5/8)^{\frac{1}{2}}.$$

Comparing (6.5) and (6.6) we obtain $\pm u_n(\pm x; A_n)$ as the only eligible extremum polynomials [and $x_0 = \pm 1$ as the points at which the extremum is obtained] provided $n \geq 5$.

7. Proof of Theorem 2 (continued)

The previous result holds also for $n = 4$, as a direct discussion shows; however, it fails for $n = 3$.

1. We have for $n = 4$:

$$(7.1) \quad T_4(x) = 8x^4 - 8x^2 + 1, \quad T_4'(x) = 32x^3 - 16x, \quad T_4''(x) = 96x^2 - 16,$$

so that, with the same notation as before, $x_4 = 6^{-\frac{1}{2}}$ and

$$(7.2) \quad |T_4'(x_4)| = (16/3)(2/3)^{\frac{1}{2}} = 4.3546 \dots$$

On the other hand, let us denote by y_1 and y_2 , the values of x for which the relative extrema of $u_4(x; A_4)$ in $-1 \leq x \leq +1$ are attained; thus $-1 < y_1 < y_2 < +1$, say. Then

$$(7.3) \quad u_4(x; A_4) = 1 - \lambda(1 - x)(B_4 - x)(y_1 - x)^2$$

must satisfy the following conditions:

$$(7.4) \quad \begin{aligned} (\alpha): \quad & u_4(-1; A_4) = -1, & \lambda(B_4 + 1)(y_1 + 1)^2 &= 1, \\ (\beta): \quad & u_4(y_2; A_4) = -1, & \lambda(1 - y_2)(B_4 - y_2)(y_1 - y_2)^2 &= 2, \\ (\gamma): \quad & u_4'(y_2; A_4) = 0, & \frac{1}{y_2 - 1} + \frac{1}{y_2 - B_4} + \frac{2}{y_2 - y_1} &= 0, \\ (\delta): \quad & u_4''(1; A_4) = 0, & 2B_4 + y_1 &= 3. \end{aligned}$$

¹⁷ See, for instance, E. Jahnke-F. Emde, *Funktionentafeln*, 1933, p. 30.

Hence $B_4 < 2$. Let

$$(7.5) \quad \begin{aligned} 1 - y_2 &= h(B_4 - 1), & B_4 - y_2 &= (h + 1)(B_4 - 1), \\ y_2 - y_1 &= (2 - h)(B_4 - 1); \end{aligned}$$

then (γ) becomes:

$$(7.6) \quad \frac{1}{h} + \frac{1}{h+1} + \frac{2}{h-2} = 0; \quad \text{i.e., } h = (1 + (33)^{\frac{1}{2}})/8 = 0.8430 \dots$$

Further, writing $v(x) = x(x+1)(x-2)^2$, we obtain from (α) and (β)

$$(7.7) \quad v\left(\frac{2}{B_4 - 1}\right) = v(h).$$

Since $v(x) = v(h)$ has h as a double root, it can be reduced to a quadratic equation giving

$$(7.8) \quad \frac{2}{B_4 - 1} = 3/2 - h + \frac{1}{2}(10h + 5)^{\frac{1}{2}} = 2.4893 \dots$$

Now

$$(7.9) \quad u'_4(1; A_4) = \lambda(B_4 - 1)(1 - y_1)^2 = 4 \frac{\frac{2}{B_4 - 1}}{v\left(\frac{2}{B_4 - 1}\right)} = 4.7881 \dots$$

Comparison of this value with (7.2) furnishes $u_4(x; A_4)$ as the solution.

2. Finally in the case $n = 3$,

$$(7.10) \quad \begin{aligned} T_3(x) &= 4x^3 - 3x, & T'_3(x) &= 12x^2 - 3, & T''_3(x) &= 24x, \\ x_3 &= 0, & |T'_3(x_3)| &= 3. \end{aligned}$$

On the other hand,

$$(7.11) \quad u_3(x; A_3) = 1 - \lambda(1 - x^2)(B_3 - x)$$

with a relative minimum at $x = y_1$, $-1 < y_1 < +1$, satisfies the following conditions:

$$(7.12) \quad \begin{aligned} (\alpha): & u_3(y_1; A_3) = -1, & \lambda(1 - y_1^2)(B_3 - y_1) &= 2, \\ (\beta): & u'_3(y_1; A_3) = 0, & 3y_1^2 - 2B_3y_1 - 1 &= 0, \\ (\gamma): & u''_3(1; A_3) = 0, & B_3 &= 3, \end{aligned}$$

so that

$$(7.13) \quad \begin{cases} y_1 = 1 - 2 \cdot 3^{-\frac{1}{2}}, & \lambda = 3^{\frac{1}{2}}/8, \\ u_3(x; A_3) = 1 - 3^{\frac{1}{2}}(1 - x^2)(3 - x)/8, & u'_3(1; A_3) = 3^{\frac{1}{2}}/2 < 3. \end{cases}$$

This completes the proof of Theorem 2.

8. Two problems of Zolotareff

1. The previous considerations permit a very simple approach to the following interesting theorem of Zolotareff:¹⁸

THEOREM 3. Let σ be a given positive number and $f(x)$ an arbitrary polynomial of degree n of the form

$$(8.1) \quad f(x) = x^n - \sigma x^{n-1} + \dots$$

Then $\max |f(x)|$, $-1 \leq x \leq +1$, is minimized if and only if

(a) $f(x) = \text{const. } u_n(x; A)$ provided $\sigma \geq n\alpha_n$,

(b) $f(x) = 2^{1-n}(1 + \sigma/n)^n T_n \left(\frac{x - \sigma/n}{1 + \sigma/n} \right)$ provided $0 < \sigma \leq n\alpha_n$.

Here $u_n(x; A)$ denotes the polynomial (2.4); and in case (a) $A = A(\sigma)$ is a uniquely determined function which increases monotonically from 1 to $+\infty$ as σ increases from $n\alpha_n$ to $+\infty$; $\alpha_n = \tan^2 [\pi/(2n)]$.

A corresponding result holds for negative σ , obtained by replacing $f(x)$ by $(-1)^n f(-x)$. For $\sigma = 0$ the extremum is given by Tchebycheff's polynomial.

From (2.4) we obtain, for $x > C$,

$$\begin{aligned} -u_n(x; A) &= Rx^n - Sx^{n-1} + \dots \\ &= \cosh \left\{ n \int_C^x (t-A)(t-B)^{-1}(t-C)^{-1}(t^2-1)^{-1} dt \right\} \\ &= \cosh \left\{ n (\log x - \log C) \right. \\ &\quad \left. + n \int_C^\infty [(t-A)(t-B)^{-1}(t-C)^{-1}(t^2-1)^{-1} - t^{-1}] dt \right. \\ &\quad \left. - n \int_x^\infty [(t-A)(t-B)^{-1}(t-C)^{-1}(t^2-1)^{-1} - t^{-1}] dt \right\}; \end{aligned}$$

so that, as $x \rightarrow +\infty$,

$$\begin{aligned} -u_n(x; A) &= \frac{1}{2}(x/C)^n \exp \left\{ n \int_C^\infty [(t-A)(t-B)^{-1}(t-C)^{-1}(t^2-1)^{-1} - t^{-1}] dt \right. \\ &\quad \left. - n \int_x^\infty [(\frac{1}{2}(B+C) - A)t^{-2} + O(t^{-3})] dt \right\} + O(x^{-n}). \end{aligned}$$

Consequently

$$(8.2) \quad R = \frac{1}{2}C^{-n} \times \exp \left\{ n \int_C^\infty [(t-A)(t-B)^{-1}(t-C)^{-1}(t^2-1)^{-1} - t^{-1}] dt \right\} > 0,$$

$$S/R = n \{ \frac{1}{2}(B+C) - A \} > 0,$$

so that R and S are continuous functions of A .

¹⁸ Loc. cit.⁸ (a), (b), (c).

From the results of Lemma 1,

$$(8.3) \quad \left(\frac{d}{dx}\right)^n u_n(x; A) = -n!R,$$

is an increasing function of A . Let $A_1 < A_2$, and let R_1, S_1, R_2, S_2 be the corresponding values of R and S , $R_1 > R_2$. Considering $R_1^{-1}u_n(x; A_1) - R_2^{-1}u_n(x; A_2)$ at the extremum points of $u_n(x; A_2)$ in $-1 \leq x \leq +1$ we see that it cannot be of degree $n-2$, so that $S_1/R_1 \neq S_2/R_2$. Hence S/R is monotonic. Its minimum value is attained for

$$u_n(x; +1) = -T_n\left(\frac{x - \alpha_n}{1 + \alpha_n}\right),$$

so that $\min(S/R) = n\alpha_n$. Its maximum value is attained for $u_n(x; +\infty) = T_{n-1}(x)$, so that $\max(S/R) = +\infty$.

Now let $f(x)$ be a polynomial of the form (8.1), and let $\sigma \geq n\alpha_n$. Then there exists a definite polynomial $u_n(x; A)$, $A = A(\sigma) \geq 1$, for which $S/R = \sigma$ so that

$$(8.4) \quad d(x) = f(x) + R^{-1}u_n(x; A)$$

is of degree $n-2$. Let $\max |f(x)| \leq R^{-1}$, $-1 \leq x \leq +1$. Then the polynomial (8.4) is alternately ≥ 0 and ≤ 0 at the points at which $u_n(x; A) = \pm 1$. Unless $d(x) \equiv 0$ this gives $n-1$ distinct points at which $d'(x)$ is alternately > 0 and < 0 , and hence $n-2$ roots for $d'(x)$ which is impossible.

2. The argument is similar in the other case, $0 < \sigma < n\alpha_n$, since the polynomial

$$(8.5) \quad 2^{1-n}(1 + \sigma/n)^n T_n\left(\frac{x - \sigma/n}{1 + \sigma/n}\right) = x^n - \sigma x^{n-1} + \dots$$

assumes its maximum modulus $2^{1-n}(1 + \sigma/n)^n$ precisely n times in $-1 \leq x \leq +1$.

Replacing $-R^{-1}u_n(x; A)$ in (8.4) by the left-hand side of (8.5), we obtain the desired result.

3. Another theorem of Zolotareff is the following¹⁹:

THEOREM 4. Let x_0, y_0 , be arbitrary real numbers, of which $x_0 > 1$, and let $f(x)$ be an arbitrary polynomial of degree n satisfying the conditions

$$(8.6) \quad f(x) = x^n + \dots, \quad f(x_0) = y_0.$$

Then $\max |f(x)|$, $-1 \leq x \leq +1$, is a minimum if and only if $f(x)$ is one of the polynomials

$$(8.7) \quad \begin{aligned} & -R^{-1}u_n(x; A), \quad 2^{1-n}(1 + \alpha)^n T_n\left(\frac{x - \alpha}{1 + \alpha}\right), \\ & 2^{1-n}(1 + \alpha)^n T_n\left(\frac{x + \alpha}{1 + \alpha}\right), \quad (-1)^{n-1} R^{-1}u_n(-x; A). \end{aligned}$$

¹⁹ Loc. cit.³ (b), p. 27, (c), p. 371.

Here $A \geq 1$, $0 \leq \alpha \leq \alpha_n = \tan^2 [\pi/(2n)]$ are certain numbers uniquely determined by x_0 and y_0 .

The values of the polynomials (8.7) at $x = x_0$ increase

from $-\infty$ to $2^{1-n}(1 + \alpha_n)^n T_n \left(\frac{x_0 - \alpha_n}{1 + \alpha_n} \right) = \beta$
as A decreases from $+\infty$ to $+1$;

from β to $2^{1-n} T_n(x_0)$ as α decreases from α_n to 0 ,

from $2^{1-n} T_n(x_0)$ to $2^{1-n}(1 + \alpha_n)^n T_n \left(\frac{x_0 + \alpha_n}{1 + \alpha_n} \right) = \beta'$
as α increases from 0 to α_n ;

from β' to $+\infty$ as A increases from 1 to $+\infty$,

respectively. These facts determine for a given y_0 the extremum polynomial $f_0(x)$ in question. Indeed, consider the difference $f(x) - f_0(x)$ at the points in $-1 \leq x \leq +1$ at which $f_0(x) = \pm 1$, and in addition at $x = x_0$. Since this difference is alternately ≥ 0 and ≤ 0 at these $n + 1$ points, the usual argument gives $n - 1$ distinct roots for its derivative [unless $f(x) \equiv f_0(x)$], which is impossible.

4. The problem defined by the condition

$$(8.8) \quad f^{(k)}(x_0) = y_0$$

where $1 \leq k \leq n - 1$, $x_0 > 1$, and y_0 is arbitrary, can be treated in a similar manner. For $k = n - 1$ we obtain the first problem dealt with above.

9. A further application

The previous considerations furnish another property of the polynomials $u_n(x; A)$ of Zolotareff which play a role in the interesting investigations of W. Markoff [see⁴].

1. We prove the following application of Lemma 1:

THEOREM 5. Let

$$(9.1) \quad 1 > x_1 > z_1 > x_2 > z_2 > \cdots > x_{n-2} > z_{n-2} > x_{n-1} > -1$$

be the values of x characterized by the conditions

$$(9.2) \quad u_n(x_\nu; A) = 0, \quad \nu = 1, 2, \dots, n - 1,$$

$$(9.3) \quad u'_n(z_\nu; A) = 0, \quad \nu = 1, 2, \dots, n - 2;$$

then the functions $x_\nu = x_\nu(A)$ and $z_\nu = z_\nu(A)$ increase as A increases.²⁰

The roots x_ν of $u_n(x; A)$ satisfy the equation

$$(9.4) \quad \int_{x_\nu}^1 (A - t)(B - t)^{-1}(C - t)^{-1}(1 - t^2)^{-1} dt = (\nu - \tfrac{1}{2})\pi/n,$$

$$\nu = 1, 2, \dots, n - 1.$$

²⁰ Concerning x_ν , see W. Markoff, loc. cit. p. 242. The largest root $D = D(A)$ also increases, as can be concluded from the result of §2, No. 4.

We can assume that $A = A(\rho)$, $B = B(\rho)$, $C = C(\rho)$ are increasing functions of a parameter ρ , $\rho > 0$, all these functions having continuous derivatives. Then $x_\nu = x_\nu(\rho)$ and

$$\begin{aligned} & (A - x_\nu)(B - x_\nu)^{-1}(C - x_\nu)^{-1}(1 - x_\nu^2)^{-1}x'_\nu(\rho) \\ &= \int_{x_\nu}^1 \frac{d}{d\rho} \{ (A - t)(B - t)^{-1}(C - t)^{-1}(1 - t^2)^{-1} \} dt \\ &= \int_{x_\nu}^1 (B - t)^{-1}(C - t)^{-1}(1 - t^2)^{-1} dt \left\{ A'(\rho) - \frac{1}{2}B'(\rho) \frac{A - t}{B - t} - \frac{1}{2}C'(\rho) \frac{A - t}{C - t} \right\} \\ &> \left\{ A'(\rho) - \frac{1}{2}B'(\rho) \frac{A + 1}{B + 1} - \frac{1}{2}C'(\rho) \frac{A + 1}{C + 1} \right\} \int_{x_\nu}^1 (B - t)^{-1}(C - t)^{-1}(1 - t^2)^{-1} dt \\ &> 0 \end{aligned}$$

since the expression (2.9) increases with ρ .

The assertion about z_ν can be proved in a similar manner.

2. The assertion about z_ν follows also from the following general remark. Suppose the roots of an algebraic equation are real and distinct, and that they are increasing functions of a parameter; then the same holds for the roots of the derivative. Indeed, using the notation above:

$$\frac{1}{z_\nu - x_1} + \frac{1}{z_\nu - x_2} + \cdots + \frac{1}{z_\nu - x_n} = 0, \quad x_n = D.$$

[Here $x_n = D$ denotes the only root of $u_n(x; A)$ which is > 1 .] Differentiating this relation,

$$\sum_{\mu=1}^n \frac{z'_\nu - x'_\mu}{(z_\nu - x_\mu)^2} = 0,$$

so that $x'_\mu > 0$ implies $z'_\nu > 0$.

Repeated application of this argument shows that the roots of all derivatives $u_n^{(k)}(x; A)$ increase as A increases.

UNIVERSITY OF PENNSYLVANIA
STANFORD UNIVERSITY

ON ABEL'S CONVERSE THEOREM

BY GUIDO FUBINI

(Received March 20, 1942)

1. Introduction

In studying Lie's famous papers¹ on the surfaces of translation, I began to think that it was perhaps useful to consider them as the statement of a theorem which, in Lie's case, is the converse of Abel's theorem, so fundamental in the theory of Abelian integrals. Lie generalized his results to other manifolds of translation. I think it will be possible to look for a generalization in another direction; in other words, it is perhaps possible to find new theorems which may be considered as the converse of Abel's theorem. If we start from a curve of genus p , we obtain some particular manifolds of translation of $p - 1$ dimensions. This has no geometrical meaning if, for instance, $p = 1$; but the question of stating for $p = 1$ a theorem which is the converse of Abel's theorem is always interesting, at least according to my opinion. Also Poincaré occupied himself with Lie's results and looked for some generalizations of the surfaces of translation. Though my point of view is different, there are some analogies between Poincaré's problem² and the problem to which this paper is devoted. But the methods of this paper are, I think, quite new; Lie starts from a system of *two partial differential equations* with *one* unknown, whereas I start from *one ordinary differential equation* with *four* unknowns (two of which are functions of a parameter u , and two functions of an independent parameter v). And I succeed in transforming this equation into a *reduced* equation which contains only the *first* derivatives with respect to one of the parameters u, v . After this the problem is easily solved, but there are many particular cases to be considered (see the conclusion of this paper). On the other hand it is not possible to take advantage of Poincaré's methods. We do not suppose from the beginning that the functions which determine our curves are analytic. Their analyticity will be a consequence of the differential equation referred to above. But even if we supposed that they are analytic, it would not be possible to make use of Poincaré's method. Let us suppose that c_1 is a curve defined by analytic functions. I choose on it a point A_1 ; afterward I define another analytic curve c_2 by giving the coordinates of its points A_2 as analytic functions of the slope m of the line A_1A_2 joining A_2 (variable on c_2) with the point A_1 chosen on c_1 . These functions may exist in a certain region of the plane of m , which we now consider as a complex variable. If this region does not contain the slope of the line t which is tangent to c_1 at A_1 , this tangent will neither intersect c_2 , nor be

¹ See Lie: *Berichte u. die Verhandl. der Kön. Sächs. Ges. der Wissensch. zu Leipzig*, 1896, Bd. 48, p. 141. *Comptes Rendus* 1892 (1st Sem.), pp. 277, 331.

² See *Bull. de la Soc. Mathem. de France*, 1901, tome 29, p. 61. (See also *Journal de Liouville*, 1895, tome 1, p. 219.)

tangent to it; and it is possible that all the straight lines of a certain neighborhood of t (for instance, the tangents t' to c_1 at the points A'_1 of a certain neighborhood of A_1) have the same property. In this case, and in other analogous cases, it is not possible to make use of Poincaré's methods, and I have not succeeded in demonstrating directly that the preceding suppositions are absurd for the curves which we shall study in our problem. In order to state this problem clearly, I recall that the simplest form of Abel's theorem for the curve of genus 1 is the following. *For a non-rational cubic we can choose the Abelian integral of first kind in such a way that its values at three points of the cubic have a sum equal to zero if and only if the three points are collinear.* (I shall later speak of the generalizations to six points lying on the same conic etc.) I now state the following converse problem. Let us suppose we have three real arcs of curves c, γ, C defined by parametric equations in which x, y are Cartesian, or, more generally, non-homogeneous projective coordinates, and u, v, w are the parameters:

$$(c) \begin{cases} x = f_1(u), \\ y = f_2(u), \end{cases} \quad (\gamma) \begin{cases} x = \varphi_1(v), \\ y = \varphi_2(v), \end{cases} \quad (C) \begin{cases} x = F_1(w), \\ y = F_2(w), \end{cases}$$

and that three points, one on every curve, are collinear if and only if

$$(1.1) \quad u + v + w = 0.$$

Can we deduce that *the three curves belong to the same curve Γ of third degree and that the corresponding parameters are identical with the Abelian integral of first kind corresponding to this curve?* The answer will be given in this paper; it is positive, but we must take into account the cases where the cubic degenerates (into a conic and a straight line or into three straight lines) or is rational (possesses a double point or a cusp).

We could also state a *more general* problem by substituting

$$\Phi_1(u) + \Phi_2(v) + \Phi_3(w) = 0$$

for (1.1) and by supposing that Φ are continuous differentiable functions of the corresponding parameter; but, by a suitable change of parameters, we realize that the new problem is equivalent to the preceding.

2. The fundamental equation

I choose at random two of the three curves c, γ, C , for instance c, γ , and I suppose that neither the former nor the latter is the line at infinity ($x = \infty, y = \infty$). (By means of a collineation we can always satisfy this condition.) On the straight line joining a point (f_1, f_2) of c with a point (φ_1, φ_2) of γ I must find a point (F_1, F_2) of C . Its coordinates will be given by

$$(1.2) \quad F_1 = kf_1 + (1 - k)\varphi_1, \quad F_2 = kf_2 + (1 - k)\varphi_2,$$

in which k is unknown. This point belongs to C ; therefore its coordinates must be functions of $w = -(u + v)$, (see (1.1)). Therefore

$$\frac{\partial F_1}{\partial u} = \frac{\partial F_1}{\partial v}, \quad \frac{\partial F_2}{\partial u} = \frac{\partial F_2}{\partial v},$$

or

$$\begin{aligned} \left[\frac{\partial k}{\partial u} - \frac{\partial k}{\partial v} \right] [f_i(u) - \varphi_i(v)] + k[f'_i(u) + \varphi'_i(v)] &= \varphi'_i(v), \\ (i = 1, 2) \quad \left[f' = \frac{dk}{du}, \varphi' = \frac{d\varphi}{dv} \right]. \end{aligned}$$

We have obtained two linear equations in two unknowns: k and $\partial k / \partial u - \partial k / \partial v$. By solving these equations, we obtain

$$(2.2) \quad \Delta k = \varphi'_2[f_1 - \varphi_1] - \varphi'_1[f_2 - \varphi_2],$$

$$(3.2) \quad \Delta \left(\frac{\partial k}{\partial u} - \frac{\partial k}{\partial v} \right) = \varphi'_1(f'_2 + \varphi'_2) - \varphi'_2(f'_1 + \varphi'_1) = \varphi'_1 f'_2 - \varphi'_2 f'_1$$

in which

$$(4.2) \quad \Delta = (f_1 - \varphi_1)(f'_2 + \varphi'_2) - (f_2 - \varphi_2)(f'_1 + \varphi'_1).$$

By differentiating (2.2) with respect to u or to v , and by subtracting, we obtain:

$$\begin{aligned} \left(\frac{\partial \Delta}{\partial u} - \frac{\partial \Delta}{\partial v} \right) k + \Delta \left(\frac{\partial k}{\partial u} - \frac{\partial k}{\partial v} \right) &= \varphi'_2 f'_1 - \varphi'_1 f'_2 + \varphi''_1(f_2 - \varphi_2) - \varphi''_2(f_1 - \varphi_1), \\ \left(f'' = \frac{d^2 f}{du^2}; \varphi'' = \frac{d^2 \varphi}{dv^2} \right). \end{aligned}$$

By recalling (3.2) we deduce:

$$\left(\frac{\partial \Delta}{\partial u} - \frac{\partial \Delta}{\partial v} \right) k = 2(\varphi'_2 f'_1 - \varphi'_1 f'_2) + \varphi''_1(f_2 - \varphi_2) - \varphi''_2(f_1 - \varphi_1).$$

By multiplying by Δ , by comparing with (2.2), and by remarking that

$$\frac{\partial \Delta}{\partial u} - \frac{\partial \Delta}{\partial v} = (f''_2 - \varphi''_2)(f_1 - \varphi_1) - (f_2 - \varphi_2)(f''_1 - \varphi''_1)$$

we conclude that

$$\begin{aligned} \{ (f''_2 - \varphi''_2)(f_1 - \varphi_1) - (f_2 - \varphi_2)(f''_1 - \varphi''_1) \} \{ \varphi'_2(f_1 - \varphi_1) - \varphi'_1(f_2 - \varphi_2) \} \\ (5.2) \quad = \{ (f_1 - \varphi_1)(f'_2 + \varphi'_2) - (f_2 - \varphi_2)(f'_1 + \varphi'_1) \} \\ \{ 2(\varphi'_2 f'_1 - \varphi'_1 f'_2) + \varphi''_1(f_2 - \varphi_2) - \varphi''_2(f_1 - \varphi_1) \}, \end{aligned}$$

or

$$\begin{aligned} \{ \varphi'_2(f_1 - \varphi_1) - \varphi'_1(f_2 - \varphi_2) \} \{ f''_2(f_1 - \varphi_1) - f''_1(f_2 - \varphi_2) - 2(\varphi'_2 f'_1 - \varphi'_1 f'_2) \} \\ (6.2) \quad = \{ f'_2(f_1 - \varphi_1) - f'_1(f_2 - \varphi_2) \} \\ \{ \varphi''_1(f_2 - \varphi_2) - \varphi''_2(f_1 - \varphi_1) + 2(\varphi'_2 f'_1 - \varphi'_1 f'_2) \}. \end{aligned}$$

The second member can be deduced from the first by interchanging the f , φ and changing the sign. This is the *fundamental equation* of our problem. If this equation is satisfied, and if $\Delta \neq 0$, the (2.2) gives us a value of k such that the functions F_1 , F_2 determined by (1.2) are functions of $w = -(u + v)$ and the three curves c , γ , C satisfy the conditions of our problem. If $\Delta = 0$, the same method which led us to (5.2) or (6.2) proves that these equations are satisfied, which can be checked directly.

If $\Delta = 0$, we have

$$(f_1 - \varphi_1)\varphi_2' - (f_2 - \varphi_2)\varphi_1' = -\{(f_1 - \varphi_1)f_2' - (f_2 - \varphi_2)f_1'\},$$

so that (5.2) or (6.2) are equivalent to the equation

$$\{\varphi_2'(f_1 - \varphi_1) - \varphi_1'(f_2 - \varphi_2)\}\{(f_2'' - \varphi_2'')(f_1 - \varphi_1) - (f_2 - \varphi_2)(f_1'' - \varphi_1'')\} = 0$$

or

$$\{\varphi_2'(f_1 - \varphi_1) - \varphi_1'(f_2 - \varphi_2)\}\left(\frac{\partial \Delta}{\partial u} - \frac{\partial \Delta}{\partial v}\right) = 0,$$

which is a consequence of the equation $\Delta = 0$. Consequently the fundamental equations are satisfied if $\Delta = 0$; but, in this case, the (2.2) does not determine the value of k . But, if we neglect the case without any geometrical meaning (at least for our problem) that c and γ are *pieces of the same straight line*, the slope

$$m = \frac{f_2 - \varphi_2}{f_1 - \varphi_1}$$

of the line joining the points (f_1, f_2) of c and (φ_1, φ_2) of γ cannot be a constant. And the equation $\Delta = 0$ is equivalent to the equation

$$\frac{\partial m}{\partial u} = \frac{\partial m}{\partial v},$$

which proves that m is a function of $w = -(u + v)$. In the exceptional case $\Delta = 0$, the conditions of our problem are satisfied too; but the third line C is the line at infinity of the plane (x, y) . Therefore (if neither c , nor γ is the line at infinity) the *fundamental equation* gives us all the necessary and sufficient conditions for these lines c , γ .

3. Remarks on the fundamental equation

We can arrive at this equation also by another way. If three points (f_1, f_2) of c , (φ_1, φ_2) of γ and (F_1, F_2) are collinear, we can suppose that $y = mx + n$ is the equation of the line joining them. Therefore

$$(1.3) \quad f_2 = mf_1 + n, \quad (2.3) \quad \varphi_2 = m\varphi_1 + n, \quad (3.3) \quad F_2 = mF_1 + n,$$

from which we deduce

$$(4.3) \quad m = \frac{f_2 - \varphi_2}{f_1 - \varphi_1}.$$

By differentiating (1.3) with respect to v , and (2.3) by respect to u , we deduce that

$$(5.3) \quad \frac{\partial n}{\partial v} = -f_1 \frac{\partial m}{\partial v}, \quad \frac{\partial n}{\partial u} = -\varphi_1 \frac{\partial m}{\partial u}.$$

Since F_i are functions of $w = -(u + v)$, we have (if $i = 1$, or 2):

$$\frac{\partial F_i}{\partial u} = \frac{\partial F_i}{\partial v}.$$

From (3.3) and (5.3) we deduce consequently

$$(6.3) \quad \frac{\partial m}{\partial u} (F_1 - \varphi_1) = \frac{\partial m}{\partial v} (F_1 - f_1), \quad \text{or} \quad F_1 = \frac{\varphi_1 \frac{\partial m}{\partial u} - f_1 \frac{\partial m}{\partial v}}{\frac{\partial m}{\partial u} - \frac{\partial m}{\partial v}}.$$

By comparing with (1.2) we deduce that

$$k = - \frac{\frac{\partial m}{\partial v}}{\frac{\partial m}{\partial u} - \frac{\partial m}{\partial v}},$$

and this equation is identical with (2.2). The fundamental equation (6.2) states only that F_1 defined by (6.3) is a function of $w = -(u + v)$.

The same method which led us to the fundamental equation proves that this equation is invariant under the group of collineations in the plane (x, y) . In order to check this statement, we shall make use of *homogeneous* projective coordinates, by writing

$$f_1 = x = \frac{x_1}{x_3}, \quad f_2 = y = \frac{x_2}{x_3}, \quad \varphi_1 = \frac{\xi_1}{\xi_3}, \quad \varphi_2 = \frac{\xi_2}{\xi_3}.$$

I shall indicate by

$$(x\xi'\xi) \text{ or } (x''\xi x), \text{ etc.},$$

the determinants of third order, the r^{th} row of which ($r = 1, 2, 3$) is

$$x_r, \xi_r', \xi_r \text{ or } x_r'', \xi_r, x_r, \text{ etc.}$$

The fundamental equation can be written in the form

$$\begin{aligned} & \begin{vmatrix} f_1 - \varphi_1 & \varphi_1' \\ f_2 - \varphi_2 & \varphi_2' \end{vmatrix} \begin{vmatrix} f_1 - \varphi_1 & f_1'' \\ f_2 - \varphi_2 & f_2'' \end{vmatrix} + \begin{vmatrix} f_1 - \varphi_1 & f_1' \\ f_2 - \varphi_2 & f_2' \end{vmatrix} \begin{vmatrix} f_1 - \varphi_1 & \varphi_1'' \\ f_2 - \varphi_2 & \varphi_2'' \end{vmatrix} \\ & = 2 \begin{vmatrix} f_1' & \varphi_1' \\ f_2' & \varphi_2' \end{vmatrix} \begin{vmatrix} f_1 - \varphi_1 & f_1' + \varphi_1' \\ f_2 - \varphi_2 & f_2' + \varphi_2' \end{vmatrix}, \end{aligned}$$

or, by supposing at first $x_3 = \xi_3 = 1$ and consequently $x_i = f_i$, $\xi_i = \varphi_i$ ($i = 1, 2$)

$$(x\xi'\xi)(xx''\xi) + (xx'\xi)(x\xi''\xi) = 2(x'\xi\xi) \begin{vmatrix} f'_1 & \varphi'_1 & A \\ f'_2 & \varphi'_2 & B \\ 0 & 0 & 1 \end{vmatrix} + 2(x\xi'\xi) \begin{vmatrix} f'_1 & \varphi'_1 & C \\ f'_2 & \varphi'_2 & D \\ 0 & 0 & 1 \end{vmatrix},$$

in which A, B, C, D are arbitrary. By supposing $A = \varphi_1$, $B = \varphi_2$, $C = f_1$, $D = f_2$, we find

$$(7.3) \quad (x\xi'\xi)(xx''\xi) + 2(x'\xi'\xi)(\xi x'x) = (\xi''x\xi)(xx'\xi) + 2(x\xi'\xi)(x'\xi'x).$$

But it is very easy to see that if we substitute ρx_i and $\sigma \xi_i$ for x_i , ξ_i , and by supposing that ρ is a function of the only u , and σ a function of the only v , this equation does not change, because both of its members are multiplied by the same factor $\rho^3 \sigma^3$. It is consequently useless to suppose $x_3 = \xi_3 = 1$. They can be arbitrary functions of u, v , respectively.

Moreover, if I transform the x, ξ by means of the same linear homogeneous transformation,³ both the members of (7.3) are multiplied by the same factor (the square of the determinant of the transformation). Therefore the form (7.3) of the fundamental equation proves that it is invariant under the group of the collineation. The two members divided, for instance, by

$$(xx'x'')(\xi\xi'\xi'')$$

are *absolute projective invariants* of the pair of curves c, γ ; they have a geometrical meaning if the parameters u, v have a geometrical meaning (for instance, if they are the projective length of an arc of the curves). I think that nobody has noticed these invariants heretofore.

4. Reduction to the first order

It does not seem to be an easy matter to study directly the fundamental equation; but we can succeed in transforming it into an equation in which there are only derivatives of first order, for instance with respect to u . In the following lines, if Φ is a function of u, v , I shall indicate by Φ' the derivative $\partial\Phi/\partial u$. For short, I shall write

$$t_i = f_i - \varphi_i, \quad t'_i = f'_i, \quad t''_i = f''_i \quad (i = 1, 2).$$

The fundamental equation can be written in the form

$$\begin{aligned} (\varphi'_2 t_1 - \varphi'_1 t_2)(t'_2 t_1 - t'_1 t_2) - 2(\varphi'_2 t'_1 - \varphi'_1 t'_2)(t_2 t_1 - t_2 t'_1) \\ = 2(\varphi'_2 t'_1 - \varphi'_1 t'_2)(\varphi'_2 t_1 - \varphi'_1 t_2) + (t'_2 t_1 - t'_1 t_2)(\varphi''_1 t_2 - \varphi''_2 t_1). \end{aligned}$$

Dividing both members by $(\varphi'_2 t_1 - \varphi'_1 t_2)^3$ I obtain

$$\frac{d}{dt} \frac{t'_2 t_1 - t'_1 t_2}{(\varphi'_2 t_1 - \varphi'_1 t_2)^2} = 2 \frac{\varphi'_2 t'_1 - \varphi'_1 t'_2}{(\varphi'_2 t_1 - \varphi'_1 t_2)^2} + \frac{t'_2 t_1 - t_2 t'_1}{(\varphi'_2 t_1 - \varphi'_1 t_2)^2} \frac{\varphi''_1 t_2 - \varphi''_2 t_1}{\varphi'_2 t_1 - \varphi'_1 t_2}.$$

³ It is now sufficient to suppose that the coefficients of the transformation are constant.

Let A, B be any two functions of v only. It is obvious that

$$\frac{d}{du} \frac{At_1 - Bt_2}{\varphi_2' t_1 - \varphi_1' t_2} = (A\varphi_1' - B\varphi_2') \frac{t_1 t_2' - t_2 t_1'}{(\varphi_2' t_1 - \varphi_1' t_2)^2}$$

$$\frac{\varphi_1'' t_2 - \varphi_2'' t_1}{\varphi_2' t_1 - \varphi_1' t_2} = \frac{\varphi_1'' \varphi_2' - \varphi_1' \varphi_2''}{A\varphi_1' - B\varphi_2'} \frac{At_1 - Bt_2}{\varphi_2' t_1 - \varphi_1' t_2} + \frac{B\varphi_2'' - A\varphi_1''}{A\varphi_1' - B\varphi_2'},$$

so that our equation becomes

$$\frac{d^2}{du^2} \frac{At_1 - Bt_2}{\varphi_2' t_1 - \varphi_1' t_2} = -2(A\varphi_1' - B\varphi_2') \frac{d}{du} \frac{1}{\varphi_2' t_1 - \varphi_1' t_2}$$

$$+ \left(\frac{\varphi_1'' \varphi_2' - \varphi_1' \varphi_2''}{A\varphi_1' - B\varphi_2'} \frac{At_1 - Bt_2}{\varphi_2' t_1 - \varphi_1' t_2} + \frac{B\varphi_2'' - A\varphi_1''}{A\varphi_1' - B\varphi_2'} \right) \frac{d}{du} \frac{At_1 - Bt_2}{\varphi_2' t_1 - \varphi_1' t_2}.$$

By integrating with respect to u , we deduce

$$(1.4) \quad \left\{ \begin{aligned} \frac{d}{du} \frac{At_1 - Bt_2}{\varphi_2' t_1 - \varphi_1' t_2} &= -2(A\varphi_1' - B\varphi_2') \frac{1}{\varphi_2' t_1 - \varphi_1' t_2} \\ &+ \frac{B\varphi_2'' - A\varphi_1''}{A\varphi_1' - B\varphi_2'} \frac{At_1 - Bt_2}{\varphi_2' t_1 - \varphi_1' t_2} + \frac{1}{2} \frac{\varphi_1'' \varphi_2' - \varphi_1' \varphi_2''}{A\varphi_1' - B\varphi_2'} \left(\frac{At_1 - Bt_2}{\varphi_2' t_1 - \varphi_1' t_2} \right)^2 + D, \end{aligned} \right.$$

in which D is a new (unknown) function of v , whereas A, B are functions of v , arbitrarily chosen, such that

$$(2.4) \quad A\varphi_1' - B\varphi_2' \neq 0.$$

(If we change A and B , the function D also will change.)

We will give to (1.4) the name of *reduced fundamental equation*. This reduction allows us to solve our problem. If v is a point of γ at which $\varphi_1' = 0$, I may suppose $B = 1$, because we can always suppose that we consider points at which the direction cosines of the tangent to γ are determined, by stating that they are proportional to φ_1', φ_2' ; so that, if $\varphi_1' = 0$, we conclude that $\varphi_2' \neq 0$; and (2.4) is satisfied if $B = 1$.

We may also suppose $A = 0$; it is easy to check that a change of A corresponds only to a change of D , and therefore we conclude:

$$(3.4) \quad \frac{d}{du} \left(\frac{f_2 - \varphi_2}{f_1 - \varphi_1} \right) = \frac{d}{du} \left(\frac{t_2}{t_1} \right) = -2\varphi_2' \frac{1}{f_1 - \varphi_1} - \frac{\varphi_2'' f_2 - \varphi_2}{\varphi_2' f_1 - \varphi_1}$$

$$+ \frac{\varphi_1''}{2\varphi_2'} \left(\frac{f_2 - \varphi_2}{f_1 - \varphi_1} \right)^2 + D_1.$$

In this equation v is no longer a variable, because we have considered a *particular* point of γ at which $\varphi_1' = 0$; therefore $\varphi_2', \varphi_2'', \varphi_1'', D_1$ are constants in (3.4).

We could consider v variable in (3.4) only if $\varphi'_1 = 0$ is an identity, or if $x = \varphi_1 = \text{const.}$ is the equation of γ . In the same manner, we find that

$$(3.4) \text{ bis } \frac{d}{du} \left(\frac{f_1 - \varphi_1}{f_2 - \varphi_2} \right) = -2\varphi'_1 \frac{1}{f_2 - \varphi_2} - \frac{\varphi''_1}{\varphi'_1} \frac{f_1 - \varphi_1}{f_2 - \varphi_2} + \frac{\varphi''_2}{2\varphi'_1} \left(\frac{f_1 - \varphi_1}{f_2 - \varphi_2} \right)^2 + D_2$$

(if $\varphi'_2 = 0$). In the general case $\varphi'_1 \varphi'_2 \neq 0$, the equation (1.4) may be written in the form:

$$\begin{aligned} t_1 t'_2 - t_2 t'_1 = & -2(\varphi'_2 t_1 - \varphi'_1 t_2) \\ & + \frac{1}{(A\varphi'_1 - B\varphi'_2)^2} \left\{ (B\varphi''_2 - A\varphi''_1)(At_1 - Bt_2)(\varphi'_2 t_1 - \varphi'_1 t_2) \right. \\ & \left. + \frac{(\varphi''_1 \varphi'_2 - \varphi'_1 \varphi''_2)(At_1 - Bt_2)^2}{2} \right\} + \frac{D}{A\varphi'_1 - B\varphi'_2} (\varphi'_2 t_1 - \varphi'_1 t_2)^2. \end{aligned}$$

The second member must obviously be independent of A, B (but D may change if we change the functions A, B). In order to check this statement it is sufficient to remark that the equation

$$\begin{aligned} 2(B\varphi''_2 - A\varphi''_1)(At_1 - Bt_2)(\varphi'_2 t_1 - \varphi'_1 t_2) + (\varphi''_1 \varphi'_2 - \varphi'_1 \varphi''_2)(At_1 - Bt_2)^2 \\ = -(lA^2 + 2mAB + nB^2)(\varphi'_2 t_1 - \varphi'_1 t_2)^2 + (nt_1^2 + 2mt_1 t_2 + lt_2^2)(\varphi'_2 B - \varphi'_1 A)^2 \end{aligned}$$

is an *identity* if l, m, n are functions of v satisfying the two following linear equations:

$$(4.4) \quad \begin{aligned} l\varphi'^2_2 - n\varphi'^2_1 &= \varphi''_1 \varphi'_2 + \varphi''_2 \varphi'_1; & m\varphi'_1 &= \varphi''_1 - l\varphi'_2 \\ & & (\text{or } m\varphi'_2 &= -\varphi''_2 - n\varphi'_1). \end{aligned}$$

Our equation becomes consequently

$$\begin{aligned} (t_1 t'_2 - t_2 t'_1) = & -2(\varphi'_2 t_1 - \varphi'_1 t_2) + \frac{1}{2}(nt_1^2 + 2mt_1 t_2 + lt_2^2) \\ & + \left[\frac{D}{A\varphi'_1 - B\varphi'_2} - \frac{lA^2 + 2mAB + nB^2}{2(A\varphi'_1 - B\varphi'_2)^2} \right] (\varphi'_2 t_1 - \varphi'_1 t_2)^2. \end{aligned}$$

But I may suppose that l, m, n satisfy not only (4.4) but also

$$(5.4) \quad 2D(A\varphi'_1 - B\varphi'_2) = lA^2 + 2mAB + nB^2.$$

The equations (4.4) and (5.4) are a system of three linear equations in l, m, n ; and it is easy to see that one can solve these equations with respect to l, m, n if $\varphi'_1 \varphi'_2 \neq 0$ and $A\varphi'_1 - B\varphi'_2 \neq 0$. We obtain

$$t_1 t'_2 - t_2 t'_1 = -2(\varphi'_2 t_1 - \varphi'_1 t_2) + \frac{1}{2}(nt_1^2 + 2mt_1 t_2 + lt_2^2)$$

in which l, m, n are functions of v satisfying (4.4). (We can deduce nothing from (5.4), because D is unknown.) By writing

$$l\varphi'^2_2 + n\varphi'^2_1 = 4H\varphi'_1 \varphi'_2$$

we deduce:

$$\begin{aligned}\frac{1}{2}l\varphi_2'^2 &= \frac{1}{4}(\varphi_1''\varphi_2' + \varphi_2''\varphi_1') + H\varphi_1'\varphi_2', & \frac{1}{2}n\varphi_1'^2 &= -\frac{1}{4}(\varphi_1''\varphi_2' + \varphi_1'\varphi_2'') + H\varphi_1'\varphi_2', \\ \frac{1}{2}m\varphi_1'\varphi_2' &= \frac{1}{2}\varphi_1''\varphi_2' - \frac{1}{2}l\varphi_2'^2 = \frac{1}{4}(\varphi_1''\varphi_2' - \varphi_1'\varphi_2'') - H\varphi_1'\varphi_2',\end{aligned}$$

so that our equation becomes

$$\begin{aligned}t_1t_2' - t_2t_1' &= -2(\varphi_2't_1 - \varphi_1't_2) + H\left(\frac{t_1}{\varphi_1'} - \frac{t_2}{\varphi_2'}\right)^2 \varphi_1'\varphi_2' \\ &\quad + \frac{1}{4}(\varphi_1''\varphi_2' + \varphi_2''\varphi_1') \left[\left(\frac{t_2}{\varphi_2'}\right)^2 - \left(\frac{t_1}{\varphi_1'}\right)^2 \right] + 2\frac{1}{4}(\varphi_1''\varphi_2' - \varphi_1'\varphi_2'') \frac{t_1t_2'}{\varphi_1'\varphi_2'}.\end{aligned}$$

If we divide this equation by $\varphi_1'\varphi_2'$ and put

$$(6.4) \quad \frac{f_1 - \varphi_1}{\varphi_1'} = z_1, \quad \frac{f_2 - \varphi_2}{\varphi_2'} = z_2$$

we obtain

$$\begin{aligned}(7.4) \quad z_1z_2' - z_2z_1' &= 2(z_2 - z_1) + H(z_1 - z_2)^2 + \frac{1}{4}\left(\frac{\varphi_1''}{\varphi_1'} + \frac{\varphi_2''}{\varphi_2'}\right)(z_2^2 - z_1^2) \\ &\quad + \frac{1}{2}\left(\frac{\varphi_1''}{\varphi_1'} - \frac{\varphi_2''}{\varphi_2'}\right)z_1z_2.\end{aligned}$$

This equation is equivalent to the reduced fundamental equation (if we interchange the indices 1, 2, we have to change only the sign of H). If we put

$$H = D\varphi_1'\varphi_2' - \frac{1}{4}\left(\frac{\varphi_1''}{\varphi_1'} + \frac{\varphi_2''}{\varphi_2'}\right)$$

this equation becomes

$$\begin{aligned}(8.4) \quad (f_1 - \varphi_1)f_2' - (f_2 - \varphi_2)f_1' &= -2[\varphi_2'(f_1 - \varphi_1) - \varphi_1'(f_2 - \varphi_2)] \\ &\quad - \frac{1}{2}\frac{\varphi_1''\varphi_2' + \varphi_2''\varphi_1'}{(\varphi_1')^2}(f_1 - \varphi_1)^2 \\ &\quad + \frac{\varphi_1''}{\varphi_1'}(f_1 - \varphi_1)(f_2 - \varphi_2) + D\{\varphi_2'(f_1 - \varphi_1) - \varphi_1'(f_2 - \varphi_2)\}^2,\end{aligned}$$

or

$$\begin{aligned}(9.4) \quad \frac{d}{du} \frac{f_1 - \varphi_1}{\Omega} &= -\frac{2}{\Omega}\varphi_1' + \frac{\varphi_1''\varphi_2' - \varphi_1'\varphi_2''}{2\varphi_1'}\left(\frac{f_1 - \varphi_1}{\Omega}\right)^2 - \frac{\varphi_1''}{\varphi_1'}\frac{f_1 - \varphi_1}{\Omega} + D\varphi_1' \\ &\quad [\Omega = \varphi_2'(f_1 - \varphi_1) - \varphi_1'(f_2 - \varphi_2)].\end{aligned}$$

5. Remarks on the reduced fundamental equation

We could repeat our considerations by substituting the curve C for the curve c . In this manner we arrive at another equation which may be deduced

by (7.4) or (6.4) by substituting F_1, F_2, w for f_1, f_2, u , and by changing, if necessary, the function H of the function D . We can demonstrate that it is sufficient to substitute F_1, F_2, w for f_1, f_2, u , *without any change of the functions H or D* . Since $w = -(u + v)$, we deduce that

$$\frac{d}{dw} = -\frac{d}{du}$$

(if v is considered as a constant). If we write the equations

$$Z_1 = \frac{F_1 - \varphi_1}{\varphi_1'}, \quad Z_2 = \frac{F_2 - \varphi_2}{\varphi_2'}$$

[which are analogous to (6.4)], we deduce from (1.2) that

$$Z_1 = kz_1, \quad Z_2 = kz_2, \quad Z' = \frac{\partial Z}{\partial w} = -\frac{\partial}{\partial u}(kz),$$

and therefore

$$Z_1 \frac{\partial Z_2}{\partial w} - Z_2 \frac{\partial Z_1}{\partial w} = -k^2 \left(z_1 \frac{\partial z_2}{\partial u} - z_2 \frac{\partial z_1}{\partial u} \right) = -k^2 (z_1 z_2' - z_2 z_1').$$

The equation, deduced from (7.4) by substituting Z, w for z, u , will be

$$\begin{aligned} -(z_1 z_2' - z_2 z_1') &= \frac{2}{k} (z_2 - z_1) + \bar{H} (z_1 - z_2)^2 \\ &\quad + \frac{1}{4} \left(\frac{\varphi_1''}{\varphi_1'} + \frac{\varphi_2''}{\varphi_2'} \right) (z_2^2 - z_1^2) + \frac{1}{2} \left(\frac{\varphi_1''}{\varphi_1'} - \frac{\varphi_2''}{\varphi_2'} \right) z_1 z_2 \end{aligned}$$

in which \bar{H} plays, for the curve C , the same role played by H for the curve c . In order to prove that $\bar{H} = H$, it is consequently necessary to prove that

$$-(z_1 z_2' - z_2 z_1') - \frac{2}{k} (z_2 - z_1) = z_1 z_2' - z_2 z_1' - 2(z_2 - z_1),$$

or

$$z_2 z_1' - z_1 z_2' + (z_2 - z_1) = \frac{1}{k} (z_2 - z_1),$$

and this is an obvious consequence of (2.2).

The reduced equation for f_1, f_2 depends on a parameter v . In the most general case we can suppose it is possible to give to v three different values such that the three equations deduced in this way from the reduced equation are independent of each other (none of them is a consequence of the other two. (Obviously this supposition has to be demonstrated.) If it is so, by eliminating the two derivatives f_1', f_2' , we obviously obtain an algebraic equation between f_1, f_2 ; and therefore the curve c would be algebraic. But the curves c and C satisfy the same reduced equation, as we have just proved. Therefore C and c would belong to

the same algebraic curve. And since c and C are any two curves chosen from among the three curves c, C, γ we can conclude that, *some particular cases excepted, the three curves c, C, γ , are arcs of the same algebraic curves.*⁴

But it is necessary to study the question more deeply in order to find the exceptional cases and to prove that the quoted algebraic curve is a cubic.

Other remarks on the reduced fundamental equation.

If we apply a collineation, the reduced fundamental equation has to be transformed into another equation which may be deduced from the initial one by changing, if necessary, the function D or H . We shall check this statement in a case of great importance for us. Instead of writing that the coordinates x, y (of a point of c) and the coordinates ξ, η (of a point of γ) are equal to $f_1, f_2, \varphi_1, \varphi_2$ respectively, we suppose that

$$x = \frac{f_1}{f_2}, \quad y = \frac{1}{f_2}, \quad f_1 = \frac{x}{y}, \quad f_2 = \frac{1}{y}, \quad \xi = \frac{\varphi_1}{\varphi_2}, \quad \eta = \frac{1}{\varphi_2}, \quad \varphi_1 = \frac{\xi}{\eta}, \quad \varphi_2 = \frac{1}{\eta}.$$

We shall write also

$$\omega = \eta'(x - \xi) - \xi'(y - \eta),$$

and we find, without difficulty, that

$$\begin{aligned} \frac{f_1 - \varphi_1}{\Omega} &= -\eta \left(\frac{\xi'\eta - \xi\eta'}{\xi'} \frac{x - \xi}{\omega} + \frac{\xi}{\xi'} \right), \\ \frac{d}{du} \left(\frac{f_1 - \varphi_1}{\Omega} \right) &= \eta(\xi'\eta - \xi\eta') \frac{x'(y - \eta) - y'(x - \xi)}{\omega^2}, \\ \frac{\varphi_1''\varphi_2' - \varphi_1'\varphi_2''}{2\varphi_1'} &= \frac{1}{2\eta} \frac{\xi'\eta'' - \eta'\xi''}{\xi'\eta - \xi\eta'}; \quad \frac{\varphi_1''}{\varphi_1'} = \frac{\xi''\eta - \xi\eta''}{\xi'\eta - \xi\eta'} - 2\frac{\eta'}{\eta}, \\ -2\frac{\varphi_1'}{\Omega} &= 2y \frac{\xi'\eta - \xi\eta'}{\omega} - (\xi'\eta - \xi\eta') \left[\frac{2\eta}{\omega} - \frac{2}{\xi'} + 2\frac{\eta'}{\xi'} \frac{x - \xi}{\omega} \right]. \end{aligned}$$

Our equation (9.4) becomes:

$$\begin{aligned} \eta(\xi'\eta - \xi\eta') \frac{x'(y - \eta) - y'(x - \xi)}{\omega^2} &= 2(\xi'\eta - \xi\eta') \left(\frac{\eta}{\xi'} + \frac{\eta'}{\xi'} \frac{x - \xi}{\omega} \right) \\ &+ \frac{\eta}{2} \frac{\xi'\eta'' - \eta'\xi''}{\xi'\eta - \xi\eta'} \left\{ \left(\frac{\xi'\eta - \xi\eta'}{\xi'} \frac{x - \xi}{\omega} \right)^2 + 2 \frac{\xi(\xi'\eta - \xi\eta')}{\xi'^2} \frac{x - \xi}{\omega} \right\} \\ &+ \left(\frac{\xi''\eta - \xi\eta''}{\xi'\eta - \xi\eta'} - 2\frac{\eta'}{\eta} \right) \eta \frac{\xi'\eta - \xi\eta'}{\xi'} \frac{x - \xi}{\omega} + \Delta\eta(\xi'\eta - \xi\eta') \end{aligned}$$

⁴ We can arrive at the same results by starting from the initial (non-reduced) fundamental equation (6.2). We ought to suppose that we can choose *five* values of v such that the corresponding *five* equations are independent. Therefore we could eliminate the *four* derivatives f_1', f_2', f_1'', f_2'' and obtain an *algebraic* equation between f_1, f_2 . This equation would also be satisfied by F_1, F_2 . But, when one follows this course, the research is much more difficult and long.

(in which Δ depends on v only), or

$$x'(y - \eta) - y'(x - \xi) = \frac{2}{\omega} + \frac{\xi'\eta'' - \eta'\xi''}{2\xi'^2} \left(\frac{x - \xi}{\omega} \right)^2 + \frac{\xi''}{\xi'^2} \frac{x - \xi}{\omega} + \Delta(v),$$

or

$$\frac{d}{du} \left(\frac{x - \xi}{\omega} \right) = -2 \frac{\xi'}{\omega} + \frac{\xi''\eta' - \xi'\eta''}{2\xi'} \left(\frac{x - \xi}{\omega} \right)^2 - \frac{\xi''}{\xi'} \frac{x - \xi}{\omega} - \Delta(v)\xi',$$

which can be deduced from (9.4) by changing the value of D and by substituting x, y, ξ, η, ω for $f_1, f_2, \varphi_1, \varphi_2, \Omega$ respectively. Moreover ω can be deduced from Ω by substituting x, y, ξ, η for $f_1, f_2, \varphi_1, \varphi_2$ respectively. We have consequently checked that (9.4) is invariant with respect to the collineation which we have considered.

6. Two of the three curves c, γ, C are straight lines

In this case I may suppose that c, γ are straight lines and suppose moreover (by using a collineation) that they are the lines $y = 0$ and $x = 0$, so that $f_2 = \varphi_1 = 0$. The fundamental equation becomes

$$\frac{f_1''}{f_1'} - 2 \frac{f_1'}{f_1} = - \left(\frac{\varphi_2''}{\varphi_2'} - 2 \frac{\varphi_2'}{\varphi_2} \right).$$

Since the first member is independent of v , and the second of u , both of them are equal to the same constant h , so that

$$(1.6) \quad \frac{d}{du} \log \frac{f_1'}{f_1^2} = h = - \frac{d}{dv} \log \frac{\varphi_2'}{\varphi_2^2}.$$

If

$$\Delta = f_1 \varphi_2' + \varphi_2 f_1'$$

should be equal to zero, we already know that the third line C would be the line at infinity. We have therefore only to study our problem if $\Delta \neq 0$. (Moreover, by means of a collineation, we could always exclude the case that C is the line at infinity.) From (1.2) and (2.2) we deduce that

$$(2.6) \quad X = F_1 = k f_1 = \frac{1}{\frac{1}{f_1} + \frac{\varphi_2 f_1'}{\varphi_2^2 f_1^2}} \quad \text{and} \quad Y = \frac{1}{\frac{1}{\varphi_2} + \frac{\varphi_2' f_1}{\varphi_2^2 f_1'}}.$$

are the parametric equations of the third curve C . If $h = 0$ we deduce from (1.6) that

$$(3.6) \quad \frac{1}{f_1} = Au + B, \quad \frac{1}{\varphi_2} = Bv + M, \quad (A, L, B, M = \text{const.})$$

$$X = \frac{1}{A(u+v) + \frac{MA+LB}{B}}, \quad Y = \frac{1}{B(u+v) + \frac{MA+LB}{A}},$$

which are functions of $w = -(u + v)$. The third line has the equation $BY - AX = 0$ and is therefore a third straight line belonging to the pencil determined by the lines c, γ . In this case the coordinates of the pencil of every one of our curves are *rational* functions of the corresponding parameters. The three curves c, γ, C form together a cubic with a triple point (which degenerates consequently into three straight lines).

If $h \neq 0$, we obtain analogously from (1.6) and (2.6),

$$(4.6) \quad \frac{1}{f_1} = Ae^{hu} + M, \quad \frac{1}{\varphi_2} = Be^{-hv} + M,$$

$$X = \frac{1}{M - \frac{NA}{B} e^{h(u+v)}}, \quad Y = \frac{1}{N - \frac{MB}{A} e^{-h(u+v)}}$$

($A, B, M, N = \text{const.}$) The third curve satisfies the linear equation $MX + NY = 1$, and is again a straight line which *does not belong* to the pencil determined by c, γ ; the coordinates of the points of every one of our lines are *not rational* functions of the corresponding parameters: it is necessary to make use of the *exponential* function. (The same thing is true if $\Delta = 0$.) The three lines c, γ, C form a *cubic with three double points*. In any case, if *two of our curves are straight lines, the third line is also a straight line*.

7. Only one of the three curves c, γ, C is a straight line

We suppose that γ is a straight line so that, by means of a collineation I may suppose $\varphi_2 = 0$ (identically). The corresponding *reduced* fundamental equation (3.4) becomes (if we write φ for φ_1):

$$\frac{d}{du} \left(\frac{f_1 - \varphi}{f_2} \right) = -2\varphi' \frac{1}{f_2} - \frac{\varphi'' f_1 - \varphi}{\varphi' f_2} + \Phi(v)$$

[Φ a function of v only] or

$$(1.7) \quad \left(\frac{f_1}{f_2} \right)' - \varphi \left(\frac{1}{f_2} \right)' = \frac{\varphi\varphi'' - 2\varphi'^2}{\varphi'} \frac{1}{f_2} - \frac{\varphi'' f_1}{\varphi' f_2} + \Phi(v).$$

By giving to v two values v_1, v_2 such that $\varphi(v_1) \neq \varphi(v_2)$, we obtain two linear equations in $(f_1/f_2)'$ and $(1/f_2)'$. By solving them, we deduce that

$$(2.7) \quad \left(\frac{f_1}{f_2} \right)' = a \frac{f_1}{f_2} + b \frac{1}{f_2} + g, \quad \left(\frac{1}{f_2} \right)' = l \frac{f_1}{f_2} + m \frac{1}{f_2} + n,$$

($a, b, g, l, m, n = \text{const.}$) and, by substituting in (1.7) we deduce that

$$(a - \varphi l) \frac{f_1}{f_2} + (b - \varphi m) \frac{1}{f_2} + (g - \varphi n) = \frac{\varphi\varphi'' - 2\varphi'^2}{\varphi'} \frac{1}{f_2} - \frac{\varphi'' f_1}{\varphi' f_2} + \Phi(v).$$

Since *only* the curve γ is a straight line, this equation must be an identity in f_1/f_2 and $1/f_2$ and consequently

$$\varphi\varphi'' - 2\varphi'^2 = (b - \varphi m)\varphi'; \quad -\varphi'' = (a - \varphi l)\varphi'$$

from which we deduce successively

$$\begin{aligned}b - \varphi m &= -2\varphi' + \varphi(l\varphi - a), \\2\varphi' &= l\varphi^2 + (m - a)\varphi - b, \\2\varphi'' &= (2l\varphi + [m - a])\varphi' = 2(l\varphi - a)\varphi'.\end{aligned}$$

Since φ is not a constant, $m - a = -2a$, or

$$(3.7) \quad a + m = 0.$$

Our equations become:

$$(4.7) \quad \begin{aligned}\varphi' &= \frac{1}{2}l\varphi^2 + m\varphi - \frac{b}{2}, \\ \left(\frac{f_1}{f_2}\right)' &= -m\frac{f_1}{f_2} + b\frac{1}{f_2} + g, \quad \left(\frac{1}{f_2}\right)' = l\frac{f_1}{f_2} + m\frac{1}{f_2} + n.\end{aligned}$$

Let us now consider the collineation

$$\bar{x} = \frac{Ax + By + E}{Lx + My + N}, \quad \bar{y} = \frac{y}{Lx + My + N} \quad \left(\begin{matrix} A, B, \dots, N = \text{const.} \\ AN - LE \neq 0 \end{matrix} \right).$$

It transforms the line γ ($y = 0$) into itself. The point (f_1, f_2) of c is transformed into the point (\bar{f}_1, \bar{f}_2) determined by

$$(5.7) \quad \frac{\bar{f}_1}{\bar{f}_2} = A\frac{f_1}{f_2} + B + E\frac{1}{f_2}, \quad \frac{1}{\bar{f}_2} = L\frac{f_1}{f_2} + M + N\frac{1}{f_2}.$$

The new function $\bar{\varphi}$ will be given by

$$(6.7) \quad \bar{\varphi} = \frac{A\varphi + E}{L\varphi + N},$$

and therefore

$$(7.7) \quad \bar{\varphi}' = \frac{1}{2} \left(l\bar{\varphi}^2 + \bar{m}\bar{\varphi} - \frac{\bar{b}}{2} \right)$$

if

$$\begin{aligned}l &= \frac{1}{AN - LE} (lN^2 - 2mLN - bL^2), \\ \bar{m} &= \frac{1}{AN - LE} (-lEN + m[LE + AN] + bLA), \\ \bar{b} &= \frac{1}{AN - LE} (-lE^2 + 2mAE + bA^2),\end{aligned}$$

from which follows

$$l\bar{b} + \bar{m}^2 = lb + m^2.$$

We could prove the same equations by studying (5.7) instead of (6.7). If $lb + m^2 \neq 0$, we can choose the values of $N:L$ and $A:E$ such that $l = \bar{b} = 0$ even if it is necessary (as in our case) that $AN - EL \neq 0$, or $N:L \neq E:A$. If $lb + m^2 = 0$, I can choose $N:L$ such that $l = 0$; since $\bar{b}l + \bar{m}^2 = lb + m^2 = 0$, \bar{m} also will be equal to zero. Therefore, by changing notation and disregarding the dashes, I can always suppose that

$$(8.7) \quad l = b = 0, \quad m \neq 0 \quad \text{or} \quad l = m = 0, \quad b \neq 0.$$

In the latter case I have added the condition $b \neq 0$. If it were not satisfied and l, m, b were equal to zero, from (4.7) we could deduce

$$\left(\frac{f_1}{f_2}\right)' = g, \quad \left(\frac{1}{f_2}\right)' = n.$$

The line c would be a straight line $f_1/f_2 = \text{const.}$, if $g = 0$, a straight line $f_2 = \text{const.}$, if $n = 0$, or a straight line $g \cdot 1/f_2 - n f_1/f_2 = \text{const.}$, if $ng \neq 0$; which is contrary to our hypothesis that among our three curves only γ is a straight line. I have now to study the two cases (8.7). In the former I can multiply u (and consequently also v) by the same constant factor (which is of no importance for our problem), such that $m = 1$; in the latter I can analogously suppose $b = 1$. Therefore in the first case:

$$\left(\frac{f_1}{f_2}\right)' = -\frac{f_1}{f_2} + g, \quad \left(\frac{1}{f_2}\right)' = \frac{1}{f_2} + n, \quad \varphi' = \varphi;$$

in the second case:

$$\left(\frac{f_1}{f_2}\right)' = \frac{1}{f_2} + g, \quad \left(\frac{1}{f_2}\right)' = n, \quad \varphi' = -\frac{1}{2}.$$

In the first case

$$\begin{aligned} \frac{f_1}{f_2} - g &= Ae^{-u}, & \frac{1}{f_2} + n &= Be^u, & \varphi &= e^{Dv}, & (A, B, D \text{ const.}) \\ (1 + nf_2)(f_1 - gf_2) &= ABf_2^2. \end{aligned}$$

The line c is the conic

$$(1 + ny)(x - gy) = AB y^2.$$

By means of (4.2), (2.2), (1.2) we calculate the coordinates $X = F_1$, $Y = F_2$ of a point of C . We find

$$X = F_1 = \frac{-g + BDe^{u+v}}{\frac{1}{D}Ae^{-(u+v)} + n}, \quad Y = F_2 = \frac{-1}{\frac{1}{D}Ae^{-(u+v)} + n}$$

[both functions of $w = -(u + v)$], and we remark that the point (X, Y) generates the same conic c , so that the curves c, C are identical. There are *two*

distinct points of intersection of the conic $c = C$ and of the line γ ($y = 0$): the point $x = y = 0$ and the point at infinity of γ . The representation of the coordinates of the points of our lines by means of the corresponding parameter requires the *exponential* function. The conic c and the line γ form together a cubic with two double points. In the second case I will apply, for short, such a collineation that $g = 0$. (I have to substitute

$$\bar{f}_1 = \frac{f_1}{gf_2 + 1}, \quad \bar{f}_2 = \frac{f_2}{gf_2 + 1}, \quad \bar{\varphi} = \bar{\varphi}_1 = \frac{\varphi_1}{g\varphi_2 + 1} = \varphi_1 = \varphi, \\ \bar{\varphi}_2 = \varphi_2 = 0, \quad \text{for } f_1, f_2, \varphi_1 \varphi_2$$

respectively, and to neglect the dashes afterward.) We find

$$\frac{1}{\bar{f}_2} = nU, \quad \frac{f_1}{\bar{f}_2} = \frac{1}{2}nU^2 + B, \quad \varphi = -\frac{1}{2}V,$$

where $U = u + \text{const.}$, $V = v + \text{const.}$, $B = \text{const.}$ Therefore the line c generated by the point $x = f_1$, $y = f_2$ is again a conic

$$2n(xy - By^2) = 1.$$

If $W = -(U + V)$, we find that the point (X, Y) of the third line C is the point

$$X = \frac{W}{2} + \frac{B}{n} \frac{1}{W}, \quad Y = \frac{1}{n} \frac{1}{W},$$

so that it generates the same conic $c = C$. In this case *rational functions* were sufficient, but the conic c and the line γ ($y = 0$) are now *tangent to each other* (at infinity). In conclusion: *If one of the curves c , γ , C is a straight line, then either the other lines are straight lines too, or they belong to the same conic. The coordinates of their points are rational functions of their parameters if the curves are lines of the same pencil, or are a straight line and a conic tangent to each other; otherwise it is necessary to make use of the exponential function.*

8. General deductions

From now on we can suppose that none of the three lines is a straight line.

I shall now choose as origin a point O of the curve γ . This point has to satisfy the following condition. If γ and c belong to algebraic curves $\bar{\gamma}$ and \bar{c} , and these algebraic curves are not identical (or, if c , γ do not belong to the same algebraic curve), the point O must not belong to \bar{c} (is not an intersection of \bar{c} , $\bar{\gamma}$). If the curves c , γ are identical, one obviously cannot satisfy this condition, and O is an arbitrary point of γ . Afterwards I choose as y -axis ($x = 0$) the tangent to γ at O . Therefore $\varphi_1 = \varphi_2 = \varphi'_1 = 0$ at O and consequently $\varphi'_2 \neq 0$. Since γ is not a straight line I can also suppose that O is not a point of inflection of γ , so that $\varphi''_1 \neq 0$ at O . The reduced equation (2.4) reads

$$(1.8) \quad \left(\frac{f_2}{f_1}\right)' = A \left(\frac{f_2}{f_1}\right)^2 + l \frac{f_2}{f_1} + m \frac{1}{f_1} + n,$$

in which the constants A, l, m, n are the values at O of $\varphi_1''/\varphi_2', -\varphi_2''/\varphi_1', -2\varphi_2', D_1$. Therefore in our case

$$(2.8) \quad Am \neq 0.$$

I write now the reduced equation (8.4) for any other point $\Omega = (\varphi_1, \varphi_2)$ of γ . And I remark that one can deduce from (1.8) that the left-hand member of (8.4) is

$$\begin{aligned} (f_1 - \varphi_1)f_2' - (f_2 - \varphi_2)f_1' &= (f_1 - \varphi_1) \left[f_1 \left(\frac{f_2}{f_1} \right)' + \frac{f_2}{f_1} f_1' \right] - (f_2 - \varphi_2) f_1' \\ &= (f_2 - \varphi_1)(lf_2 + m + nf_1) + (f_2 - \varphi_2)Af_2 + (\varphi_2 f_1 - \varphi_1 f_2) \left(A \frac{f_2}{f_1} + \frac{f_1'}{f_1} \right). \end{aligned}$$

The right-hand member of (8.4) is a polynomial of second degree in f_1, f_2 . Consequently we deduce from (8.4) that

$$-\frac{1}{f_1^2}(\varphi_2 f_2 - \varphi_1 f_2) \left(A \frac{f_2}{f_1} + \frac{f_1'}{f_1} \right) = - \left(\varphi_2 - \varphi_1 \frac{f_2}{f_1} \right) \left[A \frac{1}{f_1} \cdot \frac{f_2}{f_1} - \left(\frac{1}{f_1} \right)' \right]$$

is a polynomial P of second degree in $f_2/f_1, 1/f_1$, the coefficients of which are functions of the (arbitrary) value of v corresponding to the (arbitrary) point Ω of γ . Therefore

$$(3.8) \quad \left(\frac{1}{f_1} \right)' = A \frac{1}{f_1} \frac{f_2}{f_1} + \frac{P \left(\frac{f_2}{f_1}, \frac{1}{f_1} \right)}{\varphi_2 - \varphi_1 \frac{f_2}{f_1}}.$$

If P were divisible by $\varphi_2 - \varphi_1 f_2/f_1$ the quotient would be a polynomial of first degree:

$$(4.8) \quad a \frac{f_2}{f_1} + b \frac{1}{f_1} + g.$$

Its coefficients a, b, g ought to be independent of v (and therefore constant) because, if it were not so, by equating the values of (4.8) corresponding to two different values of v , we should find an equation of first degree in $f_2/f_1, 1/f_1$, and the curve c would be a straight line; which is contrary to our hypotheses. Therefore in this case

$$\begin{aligned} \left(\frac{f_2}{f_1} \right)' &= A \left(\frac{f_2}{f_1} \right)^2 + l \frac{f_2}{f_1} + m \frac{1}{f_1} + n, \\ (5.8) \quad \left(\frac{1}{f_1} \right)' &= A \frac{1}{f_1} \frac{f_2}{f_1} + a \frac{f_2}{f_1} + b \frac{1}{f_1} + g, \quad (A, l, m, n, a, b, g = \text{const.}) \end{aligned}$$

or

$$\begin{aligned} -f_1' &= f_1(af_2 + gf_1) + bf_1 + Af_2, \\ (6.8) \quad -f_2' &= f_2(af_2 + gf_1) - nf_1 + (b - l)f_2 - m. \end{aligned}$$

If P is not divisible by $\varphi_2 - \varphi_1 f_2/f_1$, we remark that, by giving to v two different values v_1, v_2 , we can deduce from (3.8) that

$$(7.8) \quad \frac{P_1\left(\frac{f_2}{f_1}, \frac{1}{f_1}\right)}{\varphi_2(v_1) - \varphi_1(v_1) \frac{f_2}{f_1}} = \frac{P_2\left(\frac{f_2}{f_1}, \frac{1}{f_1}\right)}{\varphi_2(v_2) - \varphi_1(v_2) \frac{f_2}{f_1}},$$

[P_1 and P_2 are the polynomials deduced from P by supposing $v = v_1$ or $v = v_2$.]

We can deduce from our supposition that (7.8) cannot be an identity in $f_2/f_1, 1/f_1$, because the denominators are not proportional to each other. (If they were proportional, the straight lines L_i defined by

$$\varphi_2(v_i) - \varphi_1(v_i) \frac{f_2}{f_1} = 0, \quad (i = 1, 2)$$

would be the same line; but these lines are the straight lines joining the origin with the points $v = v_1$ and $v = v_2$ of γ ; these three points are not collinear, because γ is not a straight line; and therefore L_1, L_2 cannot be identical.) Therefore (7.8) is an equation satisfied by all the points of c . By means of a collineation I can suppose

$$(8.8) \quad x = \frac{1}{f_1}, \quad y = \frac{f_2}{f_1}.$$

The equation (7.8) becomes

$$(9.8) \quad yS = T$$

if

$$S = \varphi_1(v_1)P_2(y, x) - \varphi_1(v_2)P_1(y, x); \quad T = \varphi_2(v_1)P_2 - \varphi_2(v_1)P_1.$$

The S, T are polynomials in x, y of a degree not greater than two. The curve c will satisfy the equation (8.8) and the equations [deduced from (1.8) and (3.8)]

$$(10.8) \quad \begin{aligned} y' &= Ay^2 + ly + mx + n, \\ x' &= Axy + \frac{Q(y, x)}{y - \eta}. \end{aligned}$$

[Q is a polynomial in x, y of a degree not greater than 2. Its coefficients are functions of v and $\eta = \varphi_2/\varphi_1$.] We have consequently to study two systems of equations: the system (5.8) and the system (9.8) and (10.8).

9. The equations (5.8) or (6.8)

These equations demonstrate that the first member of (8.4) is equal to

$$(1.9) \quad \begin{aligned} &[\varphi_1(f_2 - \varphi_2) - \varphi_2(f_1 - \varphi_1)](af_2 + gf_1) + (f_2 - \varphi_2)(bf_1 + Af_2) \\ &+ (f_1 - \varphi_1)\{nf_1 + (l - b)f_2 + n\}. \end{aligned}$$

This quantity must be equal to the second member of (8.4). We obtain in this way an equation in f_1, f_2 (of a degree not greater than two). I demonstrate that I can suppose that this is an *identity* in f_1, f_2 . If one at least of the curves c, γ, C does not belong to a conic, I can choose this curve as curve c . Since c is neither a conic nor a straight line (according to our actual hypothesis), its points cannot satisfy an equation of a degree not greater than two; and therefore the preceding equation is an identity in f_1, f_2 . If every one of the arcs c, γ, C belongs to a conic, these three conics cannot be identical to each other, because it is not possible that three points of the same conic are collinear. Therefore at least two of these conics are different from each other; and I can suppose that c, γ belong to different conics. But, by writing that (1.9) is equal to the second member of (8.4) I obtain an equation of second degree in $x = f_1, y = f_2$, which is satisfied when we suppose that $x = \varphi_1, y = \varphi_2$.⁵ If this equation were not an identity, it would be the equation of a conic, to which the point $x = \varphi_1, y = \varphi_2$ belongs; but this point is any point whatsoever of γ ; and therefore every point of γ would lie on the same conic to which c belongs; which is contrary to our supposition. Therefore (1.9) and the second member of (8.4) are identically equal to each other. I order both according to powers of $f_1 - \varphi_1, f_2 - \varphi_2$; by comparing the coefficients of similar terms, I conclude that

$$\begin{aligned}
 (2.9) \quad D\varphi_1'^2 &= a\varphi_1 + A, & \frac{\varphi_1''}{\varphi_1} - 2D\varphi_1'\varphi_2' &= l + g\varphi_1 - a\varphi_2, \\
 & - \frac{\varphi_1''\varphi_2' + \varphi_2''\varphi_1'}{2(\varphi_1')^2} + D\varphi_2'^2 &= -g\varphi_2 + n, \\
 & - 2\varphi_2' &= -\varphi_2(a\varphi_2 + g\varphi_1) + m + n\varphi_1 + (l - b)\varphi_2, \\
 & 2\varphi_1' &= \varphi_1(a\varphi_2 + g\varphi_1) + b\varphi_1 + A\varphi_2,
 \end{aligned}$$

and, by eliminating D from the first three equations:

$$\begin{aligned}
 (3.9) \quad \varphi_1'' &= 2\varphi_2'(A + a\varphi_1) + (l + g\varphi_1 - a\varphi_2)\varphi_1', \\
 \varphi_2'' &= \varphi_2'(a\varphi_2 - g\varphi_1 - l) + 2(g\varphi_2 - n)\varphi_1'.
 \end{aligned}$$

By comparing (3.9) with the values of φ'' deduced from (2.9) by differentiating, we obtain

$$(4.9) \quad (b - 2l + 3a\varphi_2)\varphi_1' - 3(A + a\varphi_1)\varphi_2' = 0,$$

$$(5.9) \quad (b + l + 3g\varphi_1)\varphi_2' + 3(n - g\varphi_2)\varphi_1' = 0.$$

If $a \neq 0$, the equation (4.9) gives

$$\frac{\varphi_2 + \frac{b - 2l}{3a}}{\varphi_1 + \frac{A}{a}} = \text{const.},$$

⁵ This would be contrary to what we supposed at the beginning of §8.

and the line γ would be a straight line, which is now absurd. Therefore $a = 0$, and the equation (4.9) proves that

$$(b - 2l)\varphi_1 - 3A\varphi_2 = \text{const.}$$

Since γ is not a straight line, we deduce

$$b - 2l = 3A = 0.$$

(We do not now take into account that $A \neq 0$ [according to (2.8)].) If $g \neq 0$, we deduce from (5.9)

$$\frac{\varphi_1 + \frac{b+l}{3g}}{\varphi_2 - \frac{n}{g}} = \text{const.},$$

and we deduce, as above, that $g = 0$; and, as above, (5.9) proves that $b + l = n = 0$. Therefore $A = a = g = b = l = n = 0$; and therefore, from (2.9) we deduce $\varphi_1 = \text{const.}$, which is absurd because γ is not a straight line. Consequently we can disregard the equations (5.8) and (6.8).

10. The equations (9.8) and (10.8)

I have already supposed that none of the three lines c , γ , C is a straight line. If every one of the lines c , γ , C belongs to a conic, I have already remarked that these three conics cannot be identical, and that I am therefore allowed to suppose that the conics to which c and γ belong are not identical. In this case I am therefore allowed to choose as origin $O(\varphi_1 = \varphi_2 = 0)$ a point of γ which does not lie on the conic to which c belongs. If

$$A_{11}f_1^2 + 2A_{12}f_1f_2 + A_{22}f_2^2 + 2A_{13}f_1 + 2A_{23}f_2 + A_{33} = 0 \quad (A_{ij} = \text{const.})$$

is the equation of the conic to which c belongs, we can therefore admit that $A_{33} \neq 0$. By means of the collineation (8.8) the equation of this conic is transformed into the equation

$$(1.10) \quad T = a_{11}x^2 + 2a_{12}xy + a_{22}y^2 + 2a_{13}x + 2a_{23}y + a_{33} = 0$$

$$(a_{11} = A_{33}, \quad a_{12} = A_{23}, \text{ etc.}),$$

in which

$$(2.10) \quad a_{11} = A_{33} \neq 0.$$

The equations (10.8) can be written in the form

$$(3.10) \quad y' = Ay^2 + ly + mx + n, \quad x' = Axy + \frac{Q(y, x)}{y - \eta},$$

in which $\eta = \varphi_2/\varphi_1$, and Q is a polynomial in y, x (of a degree not greater than two), the coefficients of which are functions of v only.

By differentiating (1.10) and taking into account (3.10) we find

$$(a_{11}x + a_{12}y + a_{13}) \left(Axy + \frac{Q(y, x)}{y - \eta} \right) \\ + (a_{21}x + a_{22}y + a_{23})(Ay^2 + ly + mx + n) = 0,$$

or

$$(4.10) \quad (y - \eta) \left[Axy + \frac{a_{21}x + a_{22}y + a_{23}}{a_{11}x + a_{12}y + a_{13}} (Ay^2 + ly + mx + n) \right] \\ = -Q(x, y).$$

By giving to v two different values, we obtain two equations; if η_i, Q_i ($i = 1, 2$) are the corresponding values of η and of Q , we find by subtracting that

$$Axy + \frac{a_{21}x + a_{22}y + a_{23}}{a_{11}x + a_{12}y + a_{13}} (Ay^2 + ly + mx + n) = \frac{Q_1 - Q_2}{\eta_1 - \eta_2} = R,$$

where R is a polynomial in x, y , of a degree not greater than two, with constant coefficients. From (4.10) we can deduce that yR also is, on the curve c belonging to the conic $T = 0$, equal to such a polynomial, and therefore we can find an identity (in x, y)

$$yR = (px + qy + r)T + (\text{a polynomial of a degree not greater than two})$$

($p, q, r = \text{const.}$). By comparing the terms in x^3 , and by recalling that $a_{11} \neq 0$, we find $p = 0$; and, since the preceding equation is an identity, we deduce that

$$R = qT + (\text{a polynomial of first degree in } x, y).$$

Therefore, on the curve c (which satisfies the equation $T = 0$) R is equal to a polynomial of first degree in x, y ; its coefficients must be constant, because γ is not a straight line. And therefore we deduce that, on c ,

$$\frac{Q}{y - \eta} = -R$$

can be considered equal to a polynomial of first degree in x, y with constant coefficients. The equations (3.10) are therefore identical with equations which we have already studied; and therefore we can suppose, from now on, that *none* of the three curves c, γ, C belongs to a straight line, and that *at least one* of them does not belong to a conic.

11. The equations (9.8) and (10.8) in the general case

We can also suppose that c does not belong to a conic, and consequently the equation (9.8) is precisely of third degree. If

$$yS_1 = T_1 \quad (S_1, T_1 \text{ polynomials in } x, y)$$

is another equation of third degree, satisfied by the points of c , this equation must be identical to (9.8) up to a constant factor k so that

$$y(S_1 - kS) = T_1 - kT \quad (\text{identically}).$$

Therefore $T_1 - kT$ must be divisible by y ; if $Lx + My + N$ is the quotient ($L, M, N = \text{const.}$), we conclude that

$$S_1 = kS + Lx + My + N, \quad T_1 = kT + y(Lx + My + N).$$

Since x' is independent of v , we deduce from (10.8) that $Q/(y - \eta)$ is independent of v . We give to v two new values v_3, v_4 , and suppose that Q_3, Q_4 are the polynomials deduced from Q in this manner. We deduce that

$$Q_3[y - \eta(v_4)] = Q_4[y - \eta(v_3)]$$

is another non-identical equation satisfied by the points of c . We deduce therefore that

$$Q_3 - Q_4 = kS + Lx + My + N,$$

$$Q_3\eta(v_4) - Q_4\eta(v_3) = kT + y(Lx + My + N),$$

in which k, L, M, N depend on v_3, v_4 . Therefore, by recalling that $yS = T$ on the curve c , we find that, on the curve c :

$$Q_3 = \frac{1}{\eta(v_4) - \eta(v_3)} \{k[T - S\eta(v_3)] + [y - \eta(v_3)](Lx + My + N)\},$$

$$\frac{Q_3}{y - \eta(v_3)} = \frac{1}{\eta(v_4) - \eta(v_3)} \{kS + Lx + My + N\},$$

which must be independent of both v_3, v_4 [see (10.8)]. Therefore, since c does not belong to a conic,

$$\frac{k}{\eta(v_4) - \eta(v_3)}, \quad \frac{L}{\eta(v_4) - \eta(v_3)}, \quad \frac{M}{\eta(v_4) - \eta(v_3)}, \quad \frac{N}{\eta(v_4) - \eta(v_3)}$$

must be constant, and therefore

$$\frac{Q}{y - \eta} = \frac{Q_3}{y - \eta(v_3)} = hS + px + qy + r \quad (h, p, q, r = \text{const.}).$$

I shall write

$$\bar{S} = hS + px + qy + r, \quad \bar{T} = hT + y(px + qy + r).$$

The equations (9.8), (10.8) become:

$$y\bar{S} = \bar{T}; \quad y' = Ay^2 + ly + mx + n, \quad x' = Axy + \bar{S}$$

or, by changing notation and disregarding the dashes:

$$(1.11) \quad yS = T,$$

$$(2.11) \quad y' = Ay^2 + ly + mx + n, \quad x' = Axy + S.$$

According to our hypotheses (§8) the origin ($\varphi_1 = \varphi_2 = 0$) does not belong to the cubic $yS = T$, if the curve γ does not belong to this cubic. We have applied a projective transformation (8.8); if we make use of homogeneous projective coordinates x_i such that $x_1:x_2:x_3 = x:y:1$, this transformation is defined by

$$x_1:x_2:x_3 = 1:f_2:f_1 \quad (\text{for the points of } c),$$

and the analogous equations

$$\xi_1:\xi_2:\xi_3 = 1:\varphi_2:\varphi_1 \quad (\text{for the points of } \gamma).$$

The equation (1.11) is turned into an equation

$$(3.11) \quad x_2 \sum a_{ik} x_i x_k = x_3 \sum t_{ik} x_i x_k,$$

(a_{ik} and t_{ik} are the coefficients of S, T) which is satisfied by the point $x_2 = x_3 = 0$ (which is the transform of the former origin). Therefore the initial origin lies on our cubic; and consequently all the curve γ belongs, like c , to this cubic.

It may perhaps be interesting to prove directly that the curves c, γ , have the same tangent at the origin $O(x_2 = x_3 = 0)$. The tangent to γ at this point was the line $\varphi_1 = 0$, which is now defined by $x_3 = 0$. By writing $x_3 = 0$ in the equation (3.11) of c , I find

$$x_2(a_{11}x^2 + 2a_{12}xy + a_{22}y^2) = 0.$$

We shall have proved that the line $x_3 = 0$ is tangent also to c , if we prove that $a_{11} = 0$. By differentiating the first of the equations (1.11) we find [by taking into account the values of x', y' given by (2.11)] that, on the curve c ,

$$\begin{aligned} 0 = [2a_{11}xy + 2a_{12}y^2 + 2a_{13}y - 2(t_{11}x + t_{12}y + t_{13})](Axy + S) \\ + [a_{11}x^2 + 4a_{12}xy + 3a_{22}y^2 + 2a_{13}x + 4a_{23}y + a_{33} - 2(t_{21}x + t_{22}y + t_{23})] \\ [Ay^2 + ly + mx + n]. \end{aligned}$$

Therefore the polynomial of fourth degree, (on the right member), must be equal to the product of $yS - T$ by a polynomial p of first degree in x, y . By comparing the terms of fourth degree in the preceding polynomials and the terms of third degree in $yS - T$, we conclude easily that

$$p = 2a_{11}x + (3A + 2a_{12})y + s \quad (s = \text{const.}).$$

By comparing the coefficients of x^3 , we deduce

$$-2t_{11}a_{11} + a_{11}m = -2a_{11}t_{11},$$

and consequently $a_{11} = 0$, because $m \neq 0$ [see (2.8)]. The right-hand members of (2.11) are polynomials in x, y of a degree not greater than two.

What happens if we make a change of projective not-homogeneous coordinates? If

$$(4.11) \quad \bar{x} = \frac{Bx + Dy + F}{Px + Qy + R}, \quad \bar{y} = \frac{Lx + My + N}{Px + Qy + R}$$

(B, D, \dots, Q, R constants, the determinant of which is different from zero) are the new coordinates, we find, for instance:

$$(5.11) \quad \bar{x}' = \frac{(BQ - DP)(x'y - xy') + (RB - FP)x' + (RD - FQ)y'}{(Px + Qy + R)^2}.$$

From (2.11) we deduce that x', y' and

$$x'y + y'x = yS - x(lx + my + n) = T - x(lx + my + n),$$

and therefore also the numerators of the right members of (5.11) are, on our curve c , equal to polynomials in x, y , the degree of which is not greater than 2. But the equations (4.11) can be written in the form

$$\begin{aligned}\bar{x}' &= B \frac{x}{Px + Qy + R} + D \frac{y}{Px + Qy + R} + F \frac{1}{Px + Qy + R}, \\ \bar{y}' &= L \frac{x}{Px + Qy + R} + M \frac{y}{Px + Qy + R} + N \frac{1}{Px + Qy + R}, \\ 1 &= P \frac{x}{Px + Qy + R} + Q \frac{y}{Px + Qy + R} + R \frac{1}{Px + Qy + R},\end{aligned}$$

and we can deduce from these equations that

$$\frac{x}{Px + Qy + R}, \quad \frac{y}{Px + Qy + R}, \quad \frac{1}{Px + Qy + R}$$

are linear integral functions of \bar{x}, \bar{y} . Therefore the right-hand member of (5.11) is a polynomial in \bar{x}, \bar{y} of a degree not higher than two; and we can obtain an analogous result for \bar{y}' . Therefore: *For any system of not-homogeneous projective coordinates x, y , their derivatives x', y' are (on c) equal to polynomials in x, y of a degree not higher than two.*

12. The canonical equation of the cubic

I can now choose such not-homogeneous projective coordinates that the equation of the cubic is

$$(1.12) \quad y^2 = 4x^3 + 2gx + r \quad (g, r = \text{const.})$$

(according to Weierstrass, $2g = -g_2, r = -g_3$). If

$$(2.12) \quad 8g^3 + 27r^2 \neq 0,$$

the cubic is not rational and has no double point. According to what we have already proved, we can write

$$(3.12) \quad x' = P_2 + P_1 + P_0, \quad y' = Q_2 + Q_1 + Q_0,$$

(P_s and Q_s are homogeneous polynomials of degree s in x, y) ($s = 0, 1, 2$). By differentiating (1.12) we deduce that

$$(4.12) \quad (6x^2 + g)(P_2 + P_1 + P_0) - y(Q_2 + Q_1 + Q_0)$$

must be equal to zero at the points of the cubic, and that therefore the polynomial (4.12) must be identically equal to the product

$$(5.12) \quad (ax + by + k)(4x^3 + 2gx + r - y^2)$$

if we choose the constants a, b, k in a suitable manner. If I compare in (4.12) and (5.12) the terms of fourth and third degree, I obtain:

$$(6.12) \quad P_2 = \frac{2}{3}x(ax + by), \quad 6x^2P_1 - yQ_2 = 4kx^3 - y^2(ax + by).$$

From the latter equation one deduces:

$$(7.12) \quad P_1 = \frac{2}{3}kx + py; \quad Q_2 = 6px^2 + y(ax + by) \quad (p = \text{an unknown constant}).$$

By comparing the terms of second degree in (4.12) and (5.12), one finds:

$$6x^2P_0 - yQ_1 + gP_2 = 2gx(ax + by) - cy^2.$$

The value of P_2 is given by (6.12); we deduce

$$(8.12) \quad P_0 = \frac{2}{3}ga, \quad Q_1 = ky - \frac{4}{3}gbx.$$

By comparing the other terms of (4.12) and (5.12), we obtain

$$gP_1 - yQ_0 = r(ax + by) + 2kgx, \quad gP_0 = kr,$$

and from (7.12), (8.12) we deduce:

$$(9.12) \quad Q_0 = gp - rb,$$

$$(10.12) \quad ra + \frac{4}{3}kg = 0, \quad 2g^2a = 9kr.$$

These last equations are linear in a, k ; if (2.12) is satisfied and the cubic is not rational, we deduce $a = k = 0$, and therefore

$$(11.12) \quad x' = P_2 + P_1 + P_0 = y(\frac{2}{3}bx + p),$$

$$(12.12) \quad \begin{aligned} y' &= Q_2 + Q_1 + Q_0 = 6px^2 + by^2 - \frac{4}{3}gbx + gp - rb \\ &= (6x^2 + g)(\frac{2}{3}bx + p). \end{aligned}$$

Therefore, if

$$(13.12) \quad U = \int \frac{dx}{y}$$

is the Abelian integral of first kind connected with our cubic, we deduce that

$$(14.12) \quad \frac{dU}{du} = \frac{2}{3}bx + p.$$

But the equation (8.4) is invariant under the group of collineations; I can consequently write it by substituting x, y, ξ, η for $f_1, f_2, \varphi_1, \varphi_2$ respectively. It proves that, on c

$$(x - \xi)y' - (y - \eta)y'$$

is equal to a polynomial of second degree in x, y . Since y' and x' , and consequently also $\xi y'$, $\eta x'$ are equal to such polynomials, $xy' - yx'$ also is a polynomial of second degree in x, y . From (11.12) and (12.12) we deduce that on our cubic

$$(\frac{2}{3}bx + p)(6x^3 + gx - y^2) = (\frac{2}{3}bx + p)(\frac{1}{3}y^2 - gx - r)$$

is such a polynomial; and this can be true only if $b = 0$. The equation (14.12) proves that u can also be considered as the Abelian integral of first kind (for our cubic), because this integral is defined up to a multiplicative and an additive constant.

We can now develop the same consideration for the curve γ , which is identical with c ; and we should find that the corresponding parameter v is given by

$$\frac{dV}{dv} = q \quad (q = \text{const.}).$$

[See the analogous equations (13.12) and (14.12).] If $p = q$ our theorem is completely demonstrated, because it is a consequence of Abel's theorem; the third curve is identical with c, γ ; the corresponding parameters are the same integral of first kind. Consequently we have to prove only that $p = q$. If $p = q$, our theorem is true and therefore the reduced equation (9.4) is satisfied, even, if we substitute, as before, x, y, ξ, η for $f_1, f_2, \varphi_1, \varphi_2$. If our problem could be solved also for another value of $p \neq q$, the equation (9.4) would be satisfied also if we write $(q/p) du$ for du , and if we change, if necessary, the function D . By subtracting the two equations which we deduce in this way from (9.4) we obtain that

$$\frac{d}{du} \frac{x - \xi}{\Omega} = \frac{d}{du} \frac{x - \xi}{\eta'(x - \xi) - \xi'(y - \eta)}$$

is only a function of v . Since $(x - \xi)/\Omega$ cannot be independent of u , because c is not a straight line, we would find, by integrating, that u is a linear (non-integer) function of x, y ; which is absurd. Our theorem is therefore completely proved.

13. The cubic is rational and possesses a double point

Let us now suppose that (2.12) is not satisfied and that the cubic possesses a double point $x = \rho, y = 0$. The equation of the cubic is now

$$(1.13) \quad y^2 = 4(x - \rho)^2(x + 2\rho) \quad (g = -2\rho^2; r = 8\rho^3).$$

At first we suppose $\rho \neq 0$. The equations (10.12) state only that $k = a\rho$ so that we obtain

$$\begin{aligned} x' &= \frac{2}{3}(ax + by) + \frac{2}{3}a\rho x + py - \frac{4}{3}a\rho^2, \\ y' &= 6px^2 + y(ax + by) + a\rho y + 8\rho^2bx - 6\rho^2p - 8\rho^3b. \end{aligned}$$

By writing that $xy' - yx'$ is, on the cubic, equal to a polynomial of second degree in x, y , we find that

$$6px^3 + \frac{1}{3}xy(ax + by)$$

must be equal to such a polynomial if we take into account the equation of c . Therefore $a = b = 0$,

$$(2.13) \quad x' = py, \quad y' = 6p(x^2 - \rho^2), \quad du = \frac{1}{p} \frac{dx}{y},$$

which is completely analogous to the definitions (13.12) of the integral of first kind for not rational cubics. In the same manner we shall find, for the parameter v corresponding to the curve γ

$$(3.13) \quad dv = \frac{1}{q} \frac{d\xi}{\eta} \quad (q = \text{const.}).$$

As above, it will be sufficient to demonstrate that our problem is solved by supposing $p = q$; which is no longer a consequence of Abel's theorem. By introducing a new parameter

$$m = \frac{y}{2(x - \rho)}$$

we find that

$$(4.13) \quad x = m^2 - 2\rho, \quad y = 2m(m^2 - \rho),$$

are new parametric equations of our cubic. We also prove easily that three points of the cubic, corresponding to the values m, μ, M of the new parameter, are collinear if and only if

$$(5.13) \quad Mm + M\mu + m\mu = -3\rho.$$

From (2.13), (4.13) it follows that

$$(6.13) \quad u = \frac{1}{p} \int \frac{dm}{m^2 - 3\rho}, \quad \text{or} \quad m = -\sqrt{3\rho} \frac{e^u + 1}{e^u - 1}$$

$$[U = 2p\sqrt{3\rho}u + h].$$

[The constant h is arbitrary.] In the same manner we find that the value μ of the new parameter can be obtained from the corresponding value of v by means of the equation

$$(7.13) \quad \mu = -\sqrt{3\rho} \frac{e^v + 1}{e^v - 1}, \quad [V = 2p\sqrt{3\rho}v + l], \quad (l = \text{const.}).$$

The equations (5.13), (6.13), (7.13) give

$$M = -\sqrt{3\rho} \frac{e^w + 1}{e^w - 1}$$

if $W = -(U + V)$. We have found the condition for the parameters, which was our starting point. Our theorem is proved also in this case; the three curves c, γ, C are identical. The introduction of the parameters U, V, W or u, v, w , requires, in this case, only the exponential function.

14. The cubic is rational and possesses a cusp

In this case $\rho = 0$; the equation of the cubic is

$$(1.14) \quad y^2 = 4x^3 \quad (g, r = 0).$$

The equations (10.12) in a, k are identities. From the results of §12 and from the remark that $xy' - yx'$ must be, on the cubic c , equal to a polynomial of second degree in x, y , we deduce that

$$(2.14) \quad x' = \frac{2}{3}kx + py, \quad y' = 6px^2 + ky.$$

We can write the reduced equation (7.4) by substituting x, y, ξ, η for $f_1, f_2, \varphi_1, \varphi_2$, respectively. From (2.14) we conclude (by taking into account that the equation of γ is $\eta^2 = 4\xi^3$ because γ is identical with c that the first member of (7.4) is

$$\begin{aligned} z_1 z_2' - z_2 z_1' &= \frac{x - \xi}{\xi'} \frac{y'}{\eta'} - \frac{y - \eta}{\eta'} \frac{x'}{\xi'} \\ &= \frac{1}{\xi' \eta'} \{ (x - \xi)(6px^2 + ky) - (y - \eta)(\frac{2}{3}kx + py) \} \\ &= \frac{p}{\xi' \eta'} \{ \frac{1}{2}(\eta' z_2)^2 + 2\eta \eta' z_2 - 6\xi \xi'^2 z_1^2 - 12\xi^2 \xi' z_1 \} \\ &\quad + k \frac{1}{\xi' \eta'} \{ \frac{1}{3}\xi' \eta' z_1 z_2 + \eta \xi' z_1 - \frac{2}{3}\xi \eta' z_2 \} \\ &\quad \left(z_1 = \frac{x - \xi}{\xi'}, z_2 = \frac{y - \eta}{\eta'} \right). \end{aligned}$$

By comparing the coefficients of $2z_1, 2z_2$ in the two members of (7.4), I obtain

$$\begin{aligned} 1 &= p \frac{\eta}{\xi'} - \frac{1}{3}k \frac{\xi}{\xi'}, \\ -1 &= -6p \frac{\xi^2}{\eta'} + \frac{1}{2}k \frac{\eta}{\eta'}. \end{aligned}$$

But $\eta/\xi' = 6\xi^2/\eta'$ is a consequence of the equation of γ ; therefore, by summing we obtain $k = 0$, and consequently

$$x' = py, \quad y' = 6px^2.$$

We find again

$$du = \frac{1}{p} \frac{dx}{y},$$

which is completely analogous to the definition of u in the other cases.

I introduce a new parameter $m = y/2x$ and find that

$$x = m^2, \quad y = 2m^3$$

are the new parametric equations of c , and that

$$u = -\frac{1}{pm} + h \quad (h = \text{const.}).$$

The introduction of u requires in this case only *rational* functions. As above, we have only to demonstrate that, by supposing

$$v = -\frac{1}{p\mu} + l \quad (l = \text{const.}),$$

our conditions are satisfied. Now the three points of c , corresponding to the values m, μ, M of the new parameter, are collinear if and only if

$$\frac{1}{m} + \frac{1}{\mu} + \frac{1}{M} = 0,$$

or if

$$u + v + w = 0,$$

in which we have written

$$w = -\frac{1}{pM} - (l + h).$$

Consequently our theorem is demonstrated in every case.

CONCLUSIONS

It is possible to state a theorem which may be considered as Abel's converse theorem for a cubic, but, besides the general case which leads to elliptic functions, we have to consider many particular cases:

1. The cubic degenerates into three straight lines forming a triangle,
2. The cubic degenerates into three straight lines belonging to the same pencil,
3. The cubic degenerates into a conic and a straight line, which intersect at two distinct points,
4. The cubic degenerates into a conic and a straight line, which are tangent to each other,
5. The cubic is rational and possesses a double point,
6. The cubic is rational and possesses a cusp.

In cases 1, 3, 5, the introduction of the corresponding parameters requires the *exponential* function, whereas *rational* functions are sufficient in cases 2, 4, 6.

GENERALIZATIONS

It is obvious that the preceding results can be easily generalized to curves of a degree higher than 1. For instance we can suppose we have six curves (or

small arcs of curves) c_1, c_2, \dots, c_6 ; the coordinates of a point of c_i are functions of a parameter u_i . I choose on every curve c_i a point A_i , and I suppose that these points lie on a conic, if and only if the sum of the corresponding parameters u_i is equal to zero. To study this problem I have to consider the conics which degenerate into two straight lines, and to choose three curves arbitrarily among the six curves c_i ,—for instance c_1, c_2, c_3 . I choose three *collinear* points A_1 on c_1 , A_2 on c_2 , A_3 on c_3 , and consider the points of c_4, c_5, c_6 and the corresponding parameters u_4, u_5, u_6 as variable. Our problem becomes: When does it happen that three points A_4 of c_4 , A_5 of c_5 , A_6 of c_6 are collinear, if

$$u_4 + u_5 + u_6 = -(u_1 + u_2 + u_3) = \text{const.}?$$

From our preceding results we deduce that c_4, c_5, c_6 form together a cubic; Therefore *three curves, chosen arbitrarily among the six given curves, form a cubic.* and it is very easy to discuss completely this new problem. We remark also that the problem is no more general if we suppose $\sum \Phi_i(u_i) = 0$ (Φ_i a function of u_i only). It may be reduced to the preceding problem by means of a change of parameters.

ANOTHER GENERALIZATION

Let us suppose we have four arcs of curves c_1, c_2, c_3, c_4 . We shall consider four functions u_1, u_2, u_3, u_4 of the points A_1, A_2, A_3, A_4 of the curves $c_1 \dots c_4$ respectively. It is possible to choose these parameters u_i in such a way that, when the points A_i are collinear, then

$$u_1 + u_2 + u_3 + u_4 = 0?$$

Let us suppose that the choice of these parameters is possible. We can state the question: Is this choice completely determined (up to non-essential constants)? If not, in how many ways can we choose linearly independent systems of such parameters u ? When we have *three* (linearly independent) systems of parameters u , our question is identical with Lie's problem on the surfaces of translation. But if one disregards for a moment these surfaces, it seems to me that the above-stated general problems and these generalizations also are very interesting from the point of view of one who studies Abel's theorem.

INSTITUTE FOR ADVANCED STUDY

HAUSDORFF INTEGRAL TRANSFORMATIONS

BY H. L. GARABEDIAN

(Received February 6, 1942)

1. Introduction

It is the object of this paper to study the integral transformation

$$(1.1) \quad v(x) = \int_0^x u(y) d\phi(y/x),$$

where $u(x)$ is bounded and continuous, $x \geq 0$, where $\phi(x)$ (called the *mass function* of the transformation) is a Hausdorff mass function (*vide infra*), and where the integration is in the sense of Riemann-Stieltjes. The transformation is said to be *regular* if the existence of $\lim_{x \rightarrow \infty} u(x)$ implies the existence of $\lim_{x \rightarrow \infty} v(x)$ and the equality of the limits.

A mass function $\phi(x)$ is said to be a Hausdorff mass function when

- (i) $\phi(x)$ is of bounded variation on the interval $0 \leq x < 1$,
- (ii) $\phi(x)$ is continuous at $x = 0$, and $\phi(0) = 0$,
- (iii) $\phi(1) = 1$,
- (iv) $\phi(x) = \frac{1}{2}[\phi(x-0) + \phi(x+0)]$ if $0 \leq x \leq 1$.

In what follows we shall use the symbols $(H, \phi(x))$ and $[H, \phi(x)]$ to designate integral and matrix transformations respectively involving the mass function $\phi(x)$.¹

The transformation (1.1) may be written in the form

$$(1.3) \quad v(x) = [1 - \phi(1-0)] u(x) + \int_0^x u(y) d\phi(y/x),$$

where a possible discontinuity of $\phi(x)$ at $x = 1$ has been removed from the integral in (1.1). In 1924 Silverman [1] studied the transformation (1.3) with the following restrictions on $\phi(x)$:

- (i) $\phi(x)$ is continuous, $0 \leq x \leq 1$,
- (ii) $\phi'(x)$ is continuous, $0 < x \leq 1$,
- (iii) $\phi(0) = 0$,
- (iv) $\int_0^1 |\phi'(x)| dx \leq M$.

¹ A discussion of the relationship of mass functions to matrix transformations may be found in [5].

In connection with these conditions it is understood that $\phi(1 - 0) = \phi(1)$. Silverman termed a function $\phi(x)$ satisfying the conditions (1.4) *absolutely regular*. Silverman proved that the transformation (1.3) is regular when $\phi(x)$ is absolutely regular,² and also established a condition in the form of an integral equation in order that $(H, \phi_a(x)) \supset (H, \phi_b(x))$, where $\phi_a(x)$ and $\phi_b(x)$ are absolutely regular mass functions.

In a recent paper [2] this writer clarified and interpreted the results of Silverman in the light of certain recent developments ([3] and [4]) in the field of Hausdorff matrix transformations, and thus made significant extensions of Silverman's results on inclusion and equivalence relations among Hausdorff integral transformations.

In this paper we extend Silverman's results to include a much wider class of mass functions than the absolutely regular mass functions just defined.

2. Regularity of $(H, \phi(x))$ when $\phi(x)$ is a Hausdorff mass function

This section is devoted primarily to a proof of the theorem which follows.

THEOREM 1. *A necessary and sufficient condition that the integral transformation $(H, \phi(x))$ be a regular transformation is that $\phi(x)$ satisfy the conditions (1.2).*

Let us first prove that if $\phi(x)$ satisfies the conditions (1.2) then the transformation (1.1) is a regular transformation. Assuming that $\lim_{x \rightarrow \infty} u(x) = l$, we wish to prove that $\lim_{x \rightarrow \infty} v(x) = l$.

We observe that

$$\int_0^x d\phi(y/x) = \int_0^1 d\phi(s) = \phi(1) - \phi(0) = 1.$$

Then we may write

$$v(x) - l = \int_0^x [u(y) - l] d\phi(y/x).$$

Now, choose p so large that $|u(x) - l| < \epsilon$, $x \geq p$. Hold p fixed and denote by M a number greater than $|u(x) - l|$ in $0 \leq x \leq p$. Then, for $x > p$:

$$\begin{aligned} |v(x) - l| &\leq \int_0^p |u(y) - l| |d\phi(y/x)| + \int_p^x |u(y) - l| |d\phi(y/x)| \\ &\leq M \int_0^p |d\phi(y/x)| + \epsilon \int_p^x |d\phi(y/x)|. \end{aligned}$$

We observe that

$$\int_p^x |d\phi(y/x)| \leq \int_0^x |d\phi(y/x)| = \int_0^1 |d\phi(s)| = V,$$

² We note also that Silverman required only that $u(x)$ be bounded and integrable, $0 \leq x \leq x_1$. In this paper we have to require the boundedness and continuity of $u(x)$, $x \geq 0$, owing to difficulties arising from the non-existence of the Stieltjes integral $\int_0^1 f(x) dg(x)$, in the case of common discontinuities of $f(x)$ and $g(x)$.

where V is the total variation of $\phi(x)$ in the interval $(0, 1)$. Moreover, since $\phi(x)$ is continuous at $x = 0$ we can find a number X so large that

$$\int_0^p |d\phi(y/x)| = \int_0^{p/x} |d\phi(s)| < \epsilon, \quad x > X.$$

Now, let X' be the greater of the numbers p and X . Then, we have

$$|v(x) - l| < \eta, \quad x > X',$$

where $\eta = (M + V)\epsilon$, whence

$$\lim_{x \rightarrow \infty} v(x) = l.$$

Finally, we show that the conditions (1.2) are necessary for the regularity of the transformation (1.1). Here we assume that $\lim_{x \rightarrow \infty} v(x) = l$ and obtain the conditions (1.2).

We observe first of all that the existence of the integral in (1.1) implies that $\phi(x)$ be a function of bounded variation on the interval $0 \leq x \leq 1$.

If in (1.1) we set $u(y) \equiv 1$, we get

$$v(x) = \int_0^x d\phi(y/x) = \phi(1) - \phi(0).$$

For regularity we must have $\phi(1) - \phi(0) = 1$. If we take $\phi(0) = 0$, then $\phi(1) = 1$.

Suppose now that $\phi(x)$ has a discontinuity at $x = 0$, that is, $\phi(+0) = \Delta$, $\phi(0) = 0$. Then, let us define the function

$$\psi(x) = \begin{cases} 0, & x = 0, \\ \Delta, & x > 0. \end{cases}$$

If we put $\phi^* = \phi - \psi$, then $\phi^*(0) = 0$ and $\phi(x)$ is continuous at $x = 0$. Note that $\phi^*(1) = 1 - \Delta$. Now we write

$$(2.1) \quad v(x) = (1 - \Delta) \int_0^x u(y) d\theta(y/x) + \int_0^x u(y) d\psi(y/x),$$

where we set $\theta(x) = \phi^*(x)/(1 - \Delta)$, so that $\theta(1) = 1$. We suppose now that $u(x)$ is any continuous function, $x \geq 0$, so that $\lim_{x \rightarrow \infty} u(x) = l$. Since $\theta(x)$ satisfies the conditions (1.2), the first integral in (2.1) defines a regular transformation and we have

$$\lim_{x \rightarrow \infty} v(x) = l(1 - \Delta) + \Delta \cdot u(0).$$

Now, we choose $u(0) = 0$. Hence, a necessary condition for regularity is the requirement $\Delta = 0$; in other words, $\phi(x)$ is continuous at $x = 0$.

We note that the regularity condition (1.2, iv) is in a sense superfluous since it serves merely to determine $\phi(x)$ uniquely at every point of the interval $(0, 1)$. With this remark the proof of our theorem is complete.

Now, we find it convenient to define $\theta_a(x) = \phi_a(1 - 0)$, $\theta_b(x) = \phi_b(1 - 0)$, $x \geq 1$. Then, we may write

$$w(x) = \alpha\beta u(x) + \alpha \int_0^1 u(xs) d\theta_b(s) + \beta \int_0^1 u(xs) d\theta_a(s) \\ + \int_0^1 \int_0^1 u(xy) d\theta_a(y/t) d\theta_b(t).$$

The last integral is of a type studied by Bray [8, Th. 5, p. 183] relative to a change in the order of integration. In the present situation the hypotheses of Bray's theorem are fulfilled with a comfortable margin of safety. Accordingly, we have

$$(3.4) \quad w(x) = \alpha\beta u(x) + \alpha \int_0^1 u(xs) d\theta_b(s) + \beta \int_0^1 u(xs) d\theta_a(s) \\ + \int_0^1 u(xy) d_y \int_0^1 \theta_a(y/t) d\theta_b(t).$$

Comparing (3.4) with (3.3) we have

$$(3.5) \quad \theta_c(s) = \alpha\theta_b(s) + \beta\theta_a(s) + \int_0^1 \theta_a(s/t) d\theta_b(t),$$

or

$$\theta_c(s) = \alpha\theta_b(s) + \beta\theta_a(s) + \theta_a(1)\theta_b(s) + \int_0^1 \theta_a(s/t) d\theta_b(t),$$

or

$$(3.6) \quad \theta_c(s) = \theta_b(s) + \beta\theta_a(s) + \int_0^1 \theta_a(s/t) d\theta_b(t).$$

Observe also that from (3.5) we can write

$$(3.7) \quad \phi_c(s) = \int_0^1 \phi_a(s/t) d\phi_b(t) = \phi_b(s) + \int_0^1 \phi_a(s/t) d\phi_b(t).$$

We now proceed to show that $\phi_c(x)$ satisfies the conditions (3.1). From (3.7) we have $\phi_c(1) = 1$. Now, in (3.6) we set $s = ty$ to obtain

$$\theta_c(s) = \theta_b(s) + \beta\theta_a(s) - \int_0^1 \theta_a(y) d\theta_b(s/y).$$

It follows now, from a theorem of Bray [8, Th. 3, p. 180], that $\theta_c(x)$ is continuous on the interval $0 \leq x \leq 1$ and hence that $\theta_c(0) = 0$. It results *a fortiori* that the conditions (ii) and (iii) of (3.1) are fulfilled. Finally, if we consider (3.6) and use another result of Bray [8, Th. 4, p. 181], it follows that $\theta_c(s)$ and hence $\phi_c(s)$ is of bounded variation on the interval $(0, 1)$.

Using (3.7) and integrating by parts we obtain

$$\phi_c(s) = \phi_a(s) - \int_s^1 \phi_b(t) d\phi_a(s/t).$$

If we set $s = ty$ we get

$$(3.8) \quad \phi_c(s) = \phi_a(s) + \int_s^1 \phi_b(s/y) d\phi_a(y).$$

This equation is of the same form as (3.7) except that ϕ_a and ϕ_b have been interchanged. Hence, $AB = BA$, whence the transformations A and B are permutable. This completes the proof of Theorem 2.

We are now in a position to state and prove the main theorem of this paper.

THEOREM 3. *Let $\phi_a(x)$ and $\phi_b(x)$ satisfy the conditions (3.1). If there exists a solution $\phi_c(x)$ of the type (3.1) which satisfies either of the Silverman-Schmidt equations:*

$$(3.9) \quad \begin{aligned} \phi_a(x) &= \int_0^1 \phi_c(x/t) d\phi_b(t), \\ \phi_a(x) &= \int_0^1 \phi_b(x/t) d\phi_c(t), \end{aligned}$$

then $(H, \phi_a(x)) \supset (H, \phi_b(x))$.

We wish to find sufficient conditions on the mass functions $\phi_a(x)$ and $\phi_b(x)$ in order that $(H, \phi_a(x)) \supset (H, \phi_b(x))$. Suppose that $\phi_a(x)$ and $\phi_b(x)$ satisfy the conditions (3.1). Let $u(x)$ be any function transformed by $(H, \phi_a(x))$ and $(H, \phi_b(x))$ into $v(x)$ and $w(x)$ respectively. We seek conditions under which $\lim_{x \rightarrow \infty} w(x) = \xi$ implies $\lim_{x \rightarrow \infty} v(x) = \xi$. Symbolically we have

$$\begin{aligned} w(x) &= B\{u(x)\} \rightarrow \xi, \\ v(x) &= A\{u(x)\}. \end{aligned}$$

Suppose that C is any transformation with an associated mass function of the type (3.1). Since C is regular we have

$$C\{w(x)\} = CB\{u(x)\} \rightarrow \xi.$$

If there exists a transformation C such that $A = CB$, then

$$\begin{aligned} \xi &= \lim_{x \rightarrow \infty} C\{w(x)\} = \lim_{x \rightarrow \infty} CB\{u(x)\} \\ &= \lim_{x \rightarrow \infty} A\{u(x)\} = \lim_{x \rightarrow \infty} v(x). \end{aligned}$$

In other words, if there exists a solution $\phi_c(x)$ of the type (3.1) satisfying either of the Silverman-Schmidt equations, we have $(H, \phi_a(x)) \supset (H, \phi_b(x))$.

4. Implications of Theorem 3

In the field of Hausdorff matrix transformations it is known that $[H, \phi_a(x)] \supset [H, \phi_b(x)]$, $\phi_a(x)$ and $\phi_b(x)$ being mass functions of the type (1.2), if and only if

there exists a mass function $\phi_c(x)$ of the type (1.2) satisfying the Silverman-Schmidt equations (3.9), for all except at most a countable set of values of x in the interval $0 < x < 1$ [4]. Thus, the Silverman-Schmidt equations serve as a connecting link between the theories of Hausdorff matrix and integral transformations. This relationship has already been discussed at considerable length by this writer in a paper already referred to [2]. It will suffice to note here that all of the inclusion relationships among Hausdorff matrix transformations involving mass functions of the type (1.4) were shown to be valid in the field of Hausdorff integral transformations. Since mass functions of the type (3.1) embrace a far wider class of Hausdorff mass functions than the absolutely regular mass functions of Silverman, Theorem 3 extends considerably the class of known inclusion relationships among Hausdorff integral transformations. We provide an example in illustration of the last statement.

Let us consider briefly the mass functions

$$\begin{aligned}\phi_1(x) &= \begin{cases} 2x, & 0 \leq x \leq 1/2, \\ 1, & 1/2 < x \leq 1, \end{cases} \\ \phi_2(x) &= x, \quad 0 \leq x \leq 1.\end{aligned}$$

We observe that $\phi_1(x)$ and $\phi_2(x)$ are mass functions of the type (3.1). In the field of Hausdorff matrix transformations it is easily proved that $[H, \phi_1(x)] \supset [H, \phi_2(x)]$. It follows that there exists a solution of the type (3.1) of the corresponding Silverman-Schmidt integral equations. We conclude finally that $(H, \phi_1(x)) \supset (H, \phi_2(x))$, and thus obtain a hitherto unknown inclusion relationship between Hausdorff integral transformations.

In this writer's opinion there is little hope of obtaining Theorem 3 for a much more general class of mass functions than those included in (3.1). There is some evidence available in support of this statement. Let us first observe that it is necessary to allow for a possible discontinuity of $\phi(x)$ at $x = 1$, since it is known that if $(H, \phi_a(x)) = (H, \phi_b(x))$ or $[H, \phi_a(x)] = [H, \phi_b(x)]$ then a solution $\phi_c(x)$ of the Silverman-Schmidt equations will have a discontinuity at $x = 1$ [2]. If we allow for additional discontinuities of $\phi(x)$ in the interval $0 < x \leq 1$, in special cases, at least, we should be led to contradictions. For example, let $\phi_a(x)$ and $\phi_b(x)$ be the mass functions associated with Euler methods of summation whose associated *moment sequences* [6] are $\{\delta_a^n\}$ and $\{\delta_b^n\}$, $0 < \delta_a < \delta_b < 1$. We have $[H, \phi_a(x)] \supset [H, \phi_b(x)]$, while (*vide supra*) $(H, \phi_a(x)) = (H, \phi_b(x))$, both methods of summation being equivalent to convergence. Consequently, the Silverman-Schmidt equations would lead to contradictory results in this case. Evidently, all mass functions of step-function character are associated with integral transformations equivalent to each other and to convergence. Whatever improvements in generality it would be possible to make in Theorem 2 would involve the tedious and difficult matter of reproving the theorems of Bray used in this paper for mass functions of a very special character.

BIBLIOGRAPHY

1. SILVERMAN, L. L., *The equivalence of certain regular transformations*, Trans. Am. Math. Soc., vol. 26 (1924), pp. 101-112.
2. GARABEDIAN, H. L., *A class of linear integral transformations*, Am. Jour. of Math., vol. 64 (1942), pp. 208-214.
3. HILLE, E. AND TAMARKIN, J. D., *Questions of relative inclusion in the domain of Hausdorff means*, Proc. Nat. Acad. Sci., vol. 19 (1933), pp. 573-577.
4. GARABEDIAN, H. L., HILLE, E., AND WALL, H. S., *Formulations of the Hausdorff inclusion problem*, Duke Math. Jour., vol. 8 (1941), pp. 193-213.
5. HAUSDORFF, F., *Summationsmethoden und Momentfolgen, I and II*, Math. Zeit., vol. 9 (1921), pp. 74-109, 280-299.
6. GARABEDIAN, H. L., *Hausdorff matrices*, Am. Math. Monthly, vol. 46 (1939), pp. 390-410.
7. SCHMIDT, R., *Über divergente Folgen und lineare Mittelbildungen*, Math. Zeits., vol. 22 (1925), pp. 89-152.
8. BRAY, H. E., *Elementary properties of the Stieltjes integral*, these Annals, vol. 20 (1919), pp. 177-186.

THE PROBLEM OF DIFFUSION OF WAVES

By J. HADAMARD

(Received March 26, 1942)

To the memory of MYRON MATHISON, whose premature death is a cruel loss to Science, I dedicate this treatment of the problem which he has solved so beautifully.

1

The various forms of Huygens' principle concern (as studied heretofore) phenomena governed by linear partial differential equations of the second order¹

$$(E) \quad F(u) = A^{kl} \frac{\partial^2 u}{\partial x^k \partial x^l} + \dots = \begin{cases} 0 & \text{(homogeneous equation)} \\ f(x) & \text{(inhomogeneous equation),} \end{cases}$$

where the terms replaced by dots contain the unknown u itself or its first derivatives, the A 's, the other coefficients, and f being given functions of the independent variables x . I have previously considered three of them,² two of which—the “major premise” and what can be called the “conclusion,” if Huygens' argument is considered as a syllogism—are general properties of such a class of phenomena. On the contrary, the “minor premise” is only true for quite special equations of the type (E): it means that when a disturbance originally located within a determinate finite region of space, propagates by waves and reaches any given point outside that region after a certain time, no effect persists *after* the passing of the wave. This is what Mathisson described by saying that these waves are *pure*.

Now, waves are not pure for any equation in an odd number of independent variables, (i.e. for wave motions in spaces with an even number of dimensions), nor are they pure when the number of variables is two. On the other hand, as is well known, the classical equation of spherical waves

$$(E_0) \quad \frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} - \frac{\partial^2 u}{\partial z^2} = 0$$

gives rise to pure waves, and this is also the case for its analogues in 6, 8, \dots , variables. The question is whether these cases are the only ones. Of course,

¹ We shall not use, at least in the present paper, the methods of the absolute differential calculus. However, we shall speak of contravariant and covariant vectors, the components being denoted by superscripts in the first case, subscripts in the second.

As usual in the absolute differential calculus, we omit, when no confusion can arise, the summation signs referring to repeated superscripts and subscripts.

² See my lecture on the subject in the Bulletin de la Société Mathématique de France, vol. LII, 1925.

two equations of the type (E) are to be considered as not essentially different if they can be deduced from each other by: 1) any point transformation

$$(a) \quad \bar{x}^h = (x^1, x^2, \dots, x^m) \quad (h = 1, 2, \dots, m)$$

on the x 's; 2) the multiplication of both sides of (E) by any (non-vanishing) factor, say

$$(b) \quad \frac{1}{\lambda} F(u) = \begin{cases} 0, \\ \frac{1}{\lambda} f(x), \end{cases} \quad \lambda = \lambda(x) = e^{\mu(x)};$$

3) multiplying the unknown by any non-vanishing factor, say

$$(c) \quad F(\lambda u) = \begin{cases} 0, \\ f. \end{cases}$$

The combination of the last two, viz. replacing (E) by

$$(bc) \quad F_{\lambda}(u) = \frac{1}{\lambda} F(\lambda u) = e^{-\mu} F(u e^{\mu}) = \begin{cases} 0, \\ e^{-\mu} f(x) \end{cases}$$

permits an easier treatment, on account of the fact that it does not change the terms of the second order and, therefore, does not alter the "characteristic form"

$$\mathbf{A}(p_1, p_2, \dots, p_m; x^1, x^2, \dots, x^m) = A^{kl} p_k p_l$$

nor the corresponding "metric form"

$$\mathfrak{A}(dx^1, dx^2, \dots, dx^m; x^1, x^2, \dots, x^m) = A_{kl} dx^k dx^l.$$

This combination is the only one which we shall deal with in this first paper.

The relations between the above two forms, which are reciprocal ones, are well known. To obtain A_{kl} , we have to divide by Δ ($\Delta = 1/D$ being the discriminant of \mathbf{A} ; D , the discriminant of \mathfrak{A}) the coefficient of A^{kl} in Δ . We have

$$A^{kh} A_{kj} = \begin{cases} 0 & (h \neq j), \\ 1 & (h = j). \end{cases}$$

The differentials dx , or, to deal with finite quantities, the m derivatives \dot{x} are to be taken as the components of a contravariant vector, the covariant components of which are the p 's if we have

$$\dot{x}^h = A^{hk} p_k,$$

or the equivalent equations

$$p_k = A_{hk} \dot{x}^h.$$

This also implies

$$\mathbf{A}(p) = \mathfrak{A}(\dot{x})$$

so that the left and right members, one being the characteristic form and the other the metric form, represent one and the same form expressed once in terms of the covariant components, then in terms of the contravariant ones.

2

Now Mathisson has succeeded in giving the above question an affirmative answer, i.e. in proving that no other simply hyperbolic³ linear partial equation in four independent variables generates pure waves, except (E_0) and those equations which are not essentially different from it. Mathisson divides his proof into two stages:

1) In the first place, he considers the case in which the coefficients A of the terms of the second order are constants;

2) In a further discussion he treats the general case.

The first part alone has been published⁴ and it will be the only one we shall deal with at present.

3

In the special case of constant A 's, we shall not have to use transformation (a), or (b) or (c) alone, but only (bc).

Furthermore, if the A 's are constant, it will always be allowed (the equation being of the simply hyperbolic type) to suppose that they have the same values as in (E_0) , viz.

$$(3.1) \quad A^{hk} = 0 \ (h \neq k), \quad A^{11} = A^{22} = A^{33} = -1, \quad A^{44} = 1.$$

However, we shall begin by not even using the first assumption; let us study the effect of (bc) on the *general* equation (E). As for the second assumption (3.1), it will be useful not to use it until the last step of the calculation.⁵

In the case of variable A 's, it is convenient to write (E) or, rather, the adjoint equation⁶

$$G(v) = 0$$

in the form⁷

$$(6) \quad \Delta_2 v + B^h \frac{\partial v}{\partial x^h} + C v = 0,$$

$$\Delta_2 v = \frac{1}{\sqrt{D}} \frac{\partial}{\partial x^h} \left(\sqrt{D} A^{hk} \frac{\partial v}{\partial x^k} \right) \text{ (second differential parameter of Lamé-Beltrami) }.$$

³ This means that the characteristic form \mathbf{A} or the equivalent metric form \mathbf{G} consists of one positive and three negative squares: what we previously called the *normal* hyperbolic case. We adopt Hilbert's and Courant's terminology.

⁴ Acta Mathematica, Vol. LXXI, 1939, pp. 249-282.

⁵ The disadvantage of immediately assuming (3.1) lies in the asymmetry of the equations between the suffixes 1, 2, 3 on one hand; 4 on the other.

⁶ In passing from (E) to its adjoint, (b) and (c) are permuted, so that (bc) remains unchanged but for the change of sign of μ .

⁷ Σ 's are omitted, as remarked in the beginning.

⁸ See, e.g., Darboux, *Leçons sur la théorie des surfaces*, vol. III.

It is known⁸ that

$$\Delta_2(\lambda v) = \lambda \Delta_2 v + 2\Delta_1(\lambda, v) + v \Delta_2 \lambda,$$

$\Delta_1(\lambda, v) = A^{hk}(\partial v / \partial x^h) \cdot (\partial \lambda / \partial x^k)$ (first parameter of Lamé-Beltrami for two functions), and that

$$e^{-\mu} \Delta_2(e^\mu) = \Delta_2 \mu + \Delta_1 \mu,$$

$\Delta_1 \mu = \Delta_1(\mu, \mu) = A^{hk}(\partial \mu / \partial x^h)(\partial \mu / \partial x^k)$, (first parameter of Lamé-Beltrami for one function) so that

$$(3.2) \quad e^{-\mu} g(v e^\mu) = \Delta_2 v + 2\bar{B}^h \frac{\partial v}{\partial x^h} + \bar{C} v,$$

$$(3.3) \quad \bar{B}^h = B^h + 2A^{hk} \frac{\partial \mu}{\partial x^k},$$

$$(3.4) \quad \bar{C} = e^{-\mu} G(e^\mu) = \Delta_2 \mu + \Delta_1 \mu + B^h \frac{\partial \mu}{\partial x^h} + C.$$

The transformation formula (3.3) on the B 's suggests substituting, in place of the quantities B^h considered as contravariant components of a vector, the corresponding covariant components

$$B_k = A_{hk} B^h$$

with respect to which the transformation formulae are simply

$$(3.3') \quad \bar{B}_k = B_k + 2 \frac{\partial \mu}{\partial x^k}.$$

We see that the differences $\bar{B}_k - B_k$ are the partial derivatives of one and the same function μ : therefore, the expressions

$$(3.5) \quad H_{hk} = \frac{\partial B_h}{\partial x^k} - \frac{\partial B_k}{\partial x^h}$$

are invariants with respect to (bc).

4

The answer to our question rests, of course, on the integration of (E) itself: more precisely, on the solution of the corresponding Cauchy problem. This integration is obtained by the use of a proper solution v of the adjoint equation, by means of which the value of the unknown u in Cauchy's problem concerning (E) is directly given by (in our case, quadruple, triple and double) integrals extended over parts of the region where the right-hand member f is defined and of the variety to which Cauchy's data refers.⁹

Instead of using the exact elementary solution, one can also get to a result

⁹ See our *Lectures on Cauchy's problem*, Cambridge-New Haven, 1923 or the French edition, Paris 1932, especially Liv. IV.

by introducing an approximate solution—what Hilbert calls a “parametrix”: then the answer to Cauchy’s problem is not directly obtained, but is reduced to an integral equation of the Fredholm (or rather Volterra) type. This is what Mathisson does. A first difficulty which he has to overcome consists, then, in showing that the necessary and sufficient condition in order that waves be pure is that the parametrix thus introduced be an exact solution, after which he has to find the cases where this condition is satisfied.

Now, it seemed to me that this roundabout use of an integral equation by introduction of a special parametrix could be avoided, since the exact elementary solution can actually be constructed; and indeed, it seems that the proof becomes simpler this way. Let us recall how the elementary solution can be calculated.¹⁰

$a(a_1, a_2, \dots, a_m)$ being an arbitrarily given point, chosen as the singular point or “pole” of the solution in question, let us consider the various geodesics issuing from a relative to the metric form \mathfrak{G} and denote by $\Gamma = \Gamma(x^1, \dots, x^m, a^1, \dots, a^m)$ the square of the geodesic distance between a and any point x : a first term (the number of independent variables still being assumed to be 4) will be V_0/Γ , with

$$(4.1) \quad V_0 = \exp \left\{ -\frac{1}{2} \int_a^x \left(\frac{M}{2} - m \right) \frac{ds}{s} \right\},$$

where

$$(4.1') \quad M = G(\Gamma) - C\Gamma = \Delta_2 \Gamma + B^h \frac{\partial \Gamma}{\partial x^h},$$

the integral being taken along the geodesic ax and s denoting the independent variable in Hamilton’s equations

$$(\mathcal{H}) \quad \dot{x}^h = \frac{dx^h}{ds} = \frac{1}{2} \frac{\partial \mathbf{A}}{\partial p_h}, \quad \dot{p}_h = \frac{dp_h}{ds} = -\frac{1}{2} \frac{\partial \mathbf{A}}{\partial x^h}$$

for that geodesic (\mathbf{A} is precisely the characteristic form, considered in Section 1).

In general, $G(V_0)$ will be different from zero; then, in order to obtain the elementary solution, the first term V_0/Γ must be completed by a term in $\log \Gamma$. The latter vanishes if, and only if

$$(4.2) \quad G(V_0) = 0$$

whenever the point x belongs to the characteristic conoid which has its vertex at a . Now, the general theory shows¹¹ that *this is the necessary and sufficient condition in order that the waves be pure.*

5

Following the same general principle as Mathisson—and we shall be able to apply it even more thoroughly than he himself has done—, we shall write our equation in a simplified form with respect to the group of the transformations

¹⁰ *Lectures on Cauchy’s problem*, Liv. II. chap. III.

¹¹ *Loc. cit.*, Liv. IV, chap. I.

(bc). Now, this is obtained at once by the consideration of V_0 : for, by an arbitrary transformation (bc), V_0 is changed into V_0/λ , as is seen by the formulae (4.1), (4.1') and (5.2) (see below): therefore, by taking $\lambda = V_0$, we get, in one and only one way, an equation for which

$$(5.1) \quad V_0 = 1,$$

a condition which, it must be remembered, *depends on the choice of the pole a*.

The above reduction condition means that the integrand in (4.1) must be identically zero in x . In the special case of constant A 's, treated in the present memoir, Γ is simply the quadratic form $\Gamma(\xi) = A_{hk}\xi^h\xi^k$; the symbol Δ_2 reduces to $A^{hk}\partial^2/\partial x^h\partial x^k$, so that, remembering (1), $\Delta_2\Gamma - 2m$ vanishes identically. As for $B^h\partial\Gamma/\partial x^h$, the derivatives of Γ being $\partial\Gamma/\partial x^h = 2P_h = 2sp_h$, it can be written, in any case, on account of (\mathcal{H}),

$$(5.2) \quad B^h \frac{\partial \Gamma}{\partial x^h} = 2sA^{hk} B_k p_k = 2sB_h x^h,$$

x^h standing for a derivative taken along the geodesic ax . When the coefficients of the terms of second order are constant, the geodesics are straight lines along which each x varies linearly, so that $sx^h = \xi^h = x^h - a^h$. In this case, therefore, $V_0 = 1$ gives us

$$(5.3) \quad B_h \xi^h = 0.$$

We shall assume that the B 's are regular in the neighborhood of a , and therefore have a Taylor's expansion

$$B_h = (B_h)_a + \xi^k \left(\frac{\partial B_h}{\partial x^k} \right)_a + \dots$$

Substituting in (5.3), we see that, in the reduced form of the equation, we must have,¹² at a ,

$$(5.4)_a \quad \left\{ \begin{array}{l} (5.4)_a \quad \bar{B}_h = 0, \\ (5.4')_a \quad \frac{\partial \bar{B}_h}{\partial x^k} + \frac{\partial \bar{B}_k}{\partial x^h} = 0, \\ (5.4'')_a \quad \frac{\partial^2 \bar{B}_h}{\partial x^k \partial x^l} + \frac{\partial^2 \bar{B}_k}{\partial x^h \partial x^l} + \frac{\partial^2 \bar{B}_l}{\partial x^h \partial x^k} = 0, \\ \dots\dots\dots \end{array} \right\} (h, k, l, \dots = 1, 2, 3, 4).$$

If, instead of taking λ equal to V_0 , we had only chosen it tangent to V_0 at a up to certain order, we should not have all the relations (5.4)_a, but only a certain (arbitrarily large) number of them. It is remarkable that Mathisson seems to have been led to write these conditions a priori, while we deduce them from (5.1), which seems to be their true origin.

¹² Throughout the following, we shall add the suffix a after the number of every relation which is proved only for $x = a$, in order to distinguish them from those which are identities in the x 's.

6

In this section, we consider the general equation (5), not assuming the A 's to be constant.

The successive relations (5.4)_a yield the successive derivatives of \bar{B}_h or, more precisely, the numerical values which they assume at a , which we shall denote by $\mu_h = (\partial\mu/\partial x^h)_a$, $\mu_{hk} = (\partial^2\mu/\partial x^h\partial x^k)_a$, \dots . The transformation formulae will be, at the point a ,

$$(6.1)_a \quad \bar{B}_h = B_h + 2\mu_h,$$

$$(6.1')_a \quad \frac{\partial \bar{B}_h}{\partial x^k} = \frac{\partial B_h}{\partial x^k} + 2\mu_{hk},$$

.....

together with the corresponding formula for \bar{C} obtained from (3.4), viz.

$$(6.2)_a \quad \bar{C} = C + B^h \mu_h + A^{hk} \mu_h \mu_k + \frac{1}{2} \frac{\partial \log |D|}{\partial x^h} A^{hk} \mu_k + \frac{\partial A^{hk}}{\partial x^h} \mu_k + A^{hk} \mu_{hk}.$$

While the set of transformations (bc) constitutes an infinite group, formulae such as the above define finite ones, since μ_h , μ_{hk} , \dots are nothing but numerical parameters: (6.1)_a are the equations of a group in m parameters; (6.1)_a, (6.1')_a and (6.2)_a, the equations of a group in $m + \frac{1}{2}m(m+1)$ parameters; and so on. In order to obtain the reduced values, we must take

$$(6.3)_a \quad \left\{ \begin{array}{l} 2\mu_h = -B_h, \\ 2\mu_{hk} = -\frac{1}{2} \left(\frac{\partial B_h}{\partial x^k} + \frac{\partial B_k}{\partial x^h} \right), \\ 2\mu_{hkl} = -\frac{1}{3} \left(\frac{\partial^2 B_h}{\partial x^k \partial x^l} + \frac{\partial^2 B_k}{\partial x^l \partial x^h} + \frac{\partial^2 B_l}{\partial x^h \partial x^k} \right), \\ \dots \end{array} \right.$$

Now, having these values of μ_h , μ_{hk} , \dots , we see, in the first place, that the reduction is made in one and only one way, so that *every equation of the class under consideration has one determinate reduced homologue with respect to the group (bc)*.

The values assumed at a by the coefficients of this reduced homologue and of their derivatives at a can be calculated in terms of the original coefficients and their derivatives, by means of (6.1), (6.1') \dots and (6.2); we find, still at a

$$(6.4)_a \quad \left\{ \begin{array}{l} \bar{B}_h = 0, \\ \frac{\partial \bar{B}_h}{\partial x^k} = \frac{1}{2} \left(\frac{\partial B_h}{\partial x^k} - \frac{\partial B_k}{\partial x^h} \right) = \frac{1}{2} H_{hk}, \\ \frac{\partial^2 \bar{B}_h}{\partial x^k \partial x^l} = \frac{1}{3} \left(\frac{\partial H_{hk}}{\partial x^l} + \frac{\partial H_{kl}}{\partial x^h} \right), \\ \dots \end{array} \right.$$

then, remembering (6.2)_a, where, on count of (6.3)_a, there is a reduction between the second and the third term, we have

$$(6.5) \quad \bar{C} = C - \frac{1}{4} B^h B_h - \frac{1}{4} \frac{\partial \log |D|}{\partial x^h} B^h - \frac{1}{2} \frac{\partial A^{hk}}{\partial x^h} B_k - \frac{1}{4} A^{hk} \left(\frac{\partial B_h}{\partial x^k} + \frac{\partial B_k}{\partial x^h} \right)$$

where¹³ the last two terms can be replaced by $-\frac{1}{2} \partial B^h / \partial x^h$.

All the quantities in the right-hand members are invariants¹⁴ of $G(v)$ with respect to the group (bc). We already knew those of the first line in (6.4)_a, and the other ones which contain only the B 's can be deduced from them by differentiation, so that they do not bring us anything new. But such is not the case as concerns the last one (6.5).

The calculations in the present section being independent¹⁴ of the hypothesis that the A 's be constant, the quantities (3.5), (6.5) are invariants, with respect to (bc), for every equation such as (E).

From now on, we come back to the case of constant A 's in which the invariant (6.5) is simplified by the vanishing of the terms in $\partial A^{hk} / \partial x^h$ or in $\partial \log |D| / \partial x^h$.

7

If we write (E) in its reduced form with respect to the pole a , so that equations (5.4)_a are satisfied, the condition $G(V_0) = 0$ is nothing else than $\bar{C} = 0$. This must happen not only at the point $x = a$, but along the whole surface of the characteristic conoid—which, in the present case, is not distinct from the characteristic cone of vertex a . In the first place we must have, at a ,

$$(7.1) \quad \bar{C} = 0$$

and this property belongs to the original (i.e. not reduced) equation—a consequence of the fact that (6.5) is an invariant. Since this must be true whatever the point a may be, we see that (7.1) *must be an identity*: we can replace C everywhere, by its value obtained from it and (6.5).

Furthermore, at a , we must have

$$(7.2)_a \quad \frac{\partial(\bar{C})}{\partial x^h} = 0,$$

$$(7.3)_a \quad \frac{\partial^2(\bar{C})}{\partial x^h \partial x^k} = q A^{hk},$$

where q is a factor common to all the terms (7.3)_a; and further successive relations expressing the fact that the function (\bar{C}) (assumed to be regular) is divisible by Γ .

¹³ This time, we do not write the suffix a , as explained in the next section.

¹⁴ For variable A 's, the equations (6.3)_a, (6.4)_a ... no longer agree with (5.1) (with the exception of the first line in (6.4)); but this is not necessary in the present section, it being essential only that the conditions define one determinate reduced homologue. We could have replaced the right hand members in (5.4) by arbitrarily given quantities Q_A, Q_{A^h}, \dots , modifying the μ_A, μ_{A^h}, \dots accordingly. It is easy to verify that we should have obtained, in that way, the same invariants with only the addition of terms containing exclusively A 's and the Q 's.

The derivatives. must be replaced
	in a derivative of \tilde{C} , by	in a derivative of (\tilde{C}) , by
$\frac{\partial \mu_k}{\partial x^h}$	$= -\frac{1}{2} \frac{\partial B_k}{\partial x^k}$	$= -\frac{1}{4} \left(\frac{\partial B_k}{\partial x^h} + \frac{\partial B_h}{\partial x^k} \right)$
$\frac{\partial^2 \mu_k}{\partial x^h \partial x^j}$	$= -\frac{1}{2} \frac{\partial^2 B_k}{\partial x^h \partial x^j}$	$= -\frac{1}{6} \left(\frac{\partial^2 B_k}{\partial x^h \partial x^j} + \frac{\partial^2 B_j}{\partial x^h \partial x^k} + \frac{\partial^2 B_h}{\partial x^j \partial x^k} \right)$
$\frac{\partial \mu_{kl}}{\partial x^h}$	$= -\frac{1}{4} \left(\frac{\partial^2 B_k}{\partial x^h \partial x^l} + \frac{\partial^2 B_l}{\partial x^h \partial x^k} \right)$	$= -\frac{1}{6} \left(\frac{\partial^2 B_k}{\partial x^h \partial x^l} + \frac{\partial^2 B_k}{\partial x^l \partial x^h} + \frac{\partial^2 B_l}{\partial x^h \partial x^k} \right)$
$\frac{\partial^2 \mu_{kl}}{\partial x^h \partial x^j}$	$= -\frac{1}{4} \left(\frac{\partial^3 B_k}{\partial x^h \partial x^j \partial x^l} + \frac{\partial^3 B_l}{\partial x^h \partial x^j \partial x^k} \right)$	$= -\frac{1}{8} \left(\frac{\partial^3 B_k}{\partial x^h \partial x^j \partial x^l} + \frac{\partial^3 B_l}{\partial x^h \partial x^j \partial x^k} + \frac{\partial^3 B_h}{\partial x^k \partial x^l \partial x^j} + \frac{\partial^3 B_j}{\partial x^k \partial x^l \partial x^h} \right)$

Since, in (7.1), \bar{C} is obtained by identifying x with a , and, thus (7.1) becomes an identity, \bar{C} and (\bar{C}) must be distinguished from each other: in (7.1), \bar{C} is the value of (6.2) obtained by replacing, at any point a (or x), μ and its derivatives by their numerical values corresponding to reduction at the same point, while, in $(7.2)_a$, $(7.3)_a$, \dots (\bar{C}) , at any point x , means the value of (3.4) when μ is constructed corresponding to reduction at a , viz. $\mu = -\log V_0(x; a)$, so that its successive derivatives, at a , are given by the formulae $(6.3)_a$.

Both quantities \bar{C} and (\bar{C}) are derived from (6.2) by substituting $-\frac{1}{2}B_h$ for μ_h and $-\frac{1}{4}(\partial B_h/\partial x^k + \partial B_k/\partial x^h)$ for μ_{hk} , so that, if not differentiated, their values are equal; another treatment is needed for their derivatives. (Cf. Table.)

Note that the corresponding values of one and the same derivative disagree only by some permutations of suffixes, so that their difference can always be expressed as a combination of the invariants H defined by (3.5).

8

Let us apply this, in the first place, to $\partial\bar{C}/\partial x^h$ and $\partial(\bar{C})/\partial x^h$, both of which we must equate to zero. We have

$$(8.1) \quad \frac{\partial\bar{C}}{\partial x^h} = \frac{\partial C}{\partial x^h} + \frac{\partial B^k}{\partial x^h} \mu_k + (B^k + 2A^{kl} \mu_l) \frac{\partial \mu_k}{\partial x^h} + A^{kl} \frac{\partial \mu^{kl}}{\partial x^h}.$$

The coefficient of $\partial \mu_k/\partial x^h$ vanishes, as μ_l is defined by the first formula $(6.3)_a$. Moreover, μ_k is the same in both expressions which we have to consider. Therefore, subtracting them from each other in order to eliminate the derivative of C , the result, simplified by interchanges of k with l in the coefficients of A^{kl} , reduces to

$$(8.2) \quad \begin{aligned} \frac{1}{6} H'_h &= A^{kl} \left[\left(\frac{\partial \mu_{kl}}{\partial x^h} \right) - \frac{\partial \mu_{kl}}{\partial x^h} \right] \\ &= \frac{1}{2} A^{kl} \left[\frac{\partial^2 B_k}{\partial x^h \partial x^l} - \frac{1}{3} \left(2 \frac{\partial^2 B_k}{\partial x^h \partial x^l} + \frac{\partial^2 B_h}{\partial x^k \partial x^l} \right) \right] = \frac{1}{6} A^{kl} \frac{\partial H_{kh}}{\partial x^l} = 0, \end{aligned}$$

which, for the same reason as (7.1), are identities.

9

These conditions are always satisfied if the H_{hk} are constant: especially, therefore, if the B 's are linear functions of the x 's. We shall now show a remarkable consequence of this, that if complex coefficients were admitted,¹⁵ waves could be pure for equations essentially distinct from (E_0) . If we do not mind introducing imaginaries, nothing prevents us from replacing, in the terms of the second order, the assumption (1) by

$$(9.1) \quad A^{hk} = 0 \quad h \neq k, \quad A^{hh} = 1 \quad (h = 1, 2, 3, 4).$$

¹⁵ Complex values could not be considered for the A 's, as the distinction between elliptic and hyperbolic cases would lose its meaning; but, theoretically, a similar objection does not hold with regard to the coefficients B .

Let us take, for each B_h , a linear polynomial

$$B_h(x) = b_h(x) + \beta_h = b_h(\xi) + B_h(a),$$

$$b_h(\xi) = \alpha_{hj} \xi^j$$

being the homogeneous part. On account of (3.2'), we can alter these expressions by the elements of any exact differential. Therefore, decomposing the matrix $\| \alpha_{hj} \|$ into a symmetric one plus an antisymmetric one, the former can be cancelled and we can assume

$$(9.2) \quad \alpha_{hj} = -\alpha_{jh},$$

and in particular

$$(9.2') \quad \alpha_{hh} = 0.$$

Then $\sum \xi^h b_h(\xi)$ vanishes and we have

$$(9.3) \quad V_0 = \exp \left\{ -\frac{1}{2} \xi^h B_h(a) \right\} = e^w.$$

On the other hand, C is given by (6.5), in which the last term disappears on account of (9.2'). Thus, substituting V_0 in G , we get

$$\begin{aligned} \frac{1}{V_0} G(V_0) &= e^{-w} G(e^w) = \frac{1}{4} \left\{ \sum [B_h(a)]^2 - 2 \sum B_h(a) B_h(x) + C \right\} \\ &= \frac{1}{4} \left\{ \sum [B_h(a)]^2 - 2 \sum B_h(a) B_h(x) + \sum [B_h(x)]^2 \right\} \\ &= \frac{1}{4} \sum [b_h(\xi)]^2. \end{aligned}$$

This must be zero on the conoid which has its vertex at a , so that the above quadratic form must be proportional to $\sum_0^3 (\xi^h)^2$, and we are led to the problem of expressing the sum $\sum (\xi^h)^2$ as a sum of squares of linear forms, the coefficients of which form an antisymmetric matrix.

As is easily seen, the most general way of obtaining this is to complete (9.2) into

$$(9.4) \quad \alpha_{hj} = -\alpha_{jh} = \epsilon \alpha_{kl} = -\epsilon \alpha_{lk}, \quad \epsilon = \pm 1,$$

where $h j k l$ is any even (or alternate) permutation of the suffixes 1, 2, 3, 4.

In other words, we have only to take $\alpha = 0$ in the well-known identity

$$\begin{aligned} (\alpha^2 + \beta^2 + \gamma^2 + \delta^2)(x^2 + y^2 + z^2 + t^2) \\ = (\alpha x + \beta y + \gamma z + \delta t)^2 + (-\beta x + \alpha y + \gamma z + \delta t)^2 \\ + (-\gamma x - \delta y + \alpha z + \beta t)^2 + (-\delta x + \gamma y - \beta z + \alpha t)^2 \end{aligned}$$

which gives the transformation of the product of two sums of four squares into a sum of four squares. A simple form of the result, to which the most general one can be reduced by an orthogonal change of the variables (for instance, an

orthogonal change of x and z combined with an orthogonal change on x and t) is the equation

$$F(u) = \Delta u + 2 \left(y \frac{\partial u}{\partial x} - x \frac{\partial u}{\partial y} + t \frac{\partial u}{\partial z} - z \frac{\partial u}{\partial t} \right) + u(x^2 + y^2 + z^2 + t^2) = 0$$

which, for any given a, b, c, d , admits of the solution

$$u = \frac{V_0}{(x-a)^2 + (y-b)^2 + (z-c)^2 + (t-d)^2} - w$$

$$= \frac{e^{ay-bx+ct-ds}}{(x-a)^2 + (y-b)^2 + (z-c)^2 + (t-d)^2} - w,$$

w being any holomorphic solution of $F(u) = V_0$.

Changing t into it , this would give a hyperbolic equation corresponding to pure waves. But we have reached this only by introduction of imaginaries. On the contrary, we shall now see that no such solution can exist in the real domain.

10

Let us treat (7.3)_a as we did for (7.2)_a. We have

$$\frac{\partial^2 \bar{C}}{\partial x^h \partial x^j} = \frac{\partial^2 C}{\partial x^h \partial x^j} + \frac{\partial^2 B^k}{\partial x^h \partial x^j} \mu_k + \left(\frac{\partial B^k}{\partial x^h} \frac{\partial \mu_k}{\partial x^j} + \frac{\partial B^k}{\partial x^j} \frac{\partial \mu_k}{\partial x^h} \right)$$

$$+ 2A^{kl} \frac{\partial \mu_l}{\partial x^h} \frac{\partial \mu_k}{\partial x^j} + (B^k + 2A^{kl} \mu_l) \frac{\partial^2 \mu_k}{\partial x^h \partial x^j} + A^{kl} \frac{\partial^2 \mu_{kl}}{\partial x^h \partial x^j}$$

As in the treatment of (7.2)_a, we remark that the coefficient of $\partial^2 \mu_k / \partial x^h \partial x^j$ vanishes and the second term is the same in $\partial^2 \bar{C} / \partial x^h \partial x^j$ and in $\partial^2 (\bar{C}) / \partial x^h \partial x^j$. As for the terms in $\partial^2 \mu_{kl} / \partial x^h \partial x^j$, they give, in $\partial^2 (\bar{C}) / \partial x^h \partial x^j - \partial^2 \bar{C} / \partial x^h \partial x^j$, the result

$$A^{kl} \left[\frac{1}{2} \frac{\partial^3 B_k}{\partial x^h \partial x^j \partial x^l} - \frac{1}{8} \left(2 \frac{\partial^3 B_k}{\partial x^h \partial x^j \partial x^l} + \frac{\partial^3 B_j}{\partial x^h \partial x^k \partial x^l} + \frac{\partial^3 B_h}{\partial x^j \partial x^k \partial x^l} \right) \right]$$

$$= \frac{1}{8} A^{kl} \left(\frac{\partial^2 H_{kh}}{\partial x^j \partial x^l} + \frac{\partial^2 H_{kj}}{\partial x^h \partial x^l} \right) = \frac{1}{8} \left(\frac{\partial H'_{kh}}{\partial x^j} + \frac{\partial H'_{kj}}{\partial x^h} \right)$$

which vanishes on account of (8.2). Moreover, we have

$$\left(\frac{\partial \mu_k}{\partial x^j} \right) - \frac{\partial \mu_k}{\partial x^j} = \frac{1}{4} H_{kj}, \quad \left(\frac{\partial \mu_k}{\partial x^h} \right) - \frac{\partial \mu_k}{\partial x^h} = \frac{1}{4} H_{kh}$$

so that we can write for the difference $\frac{\partial^2 (\bar{C})}{\partial x^h \partial x^j} - \frac{\partial^2 \bar{C}}{\partial x^h \partial x^j}$,

$$(10.1) \quad \frac{\partial^2 (\bar{C})}{\partial x^h \partial x^j} - \frac{\partial^2 \bar{C}}{\partial x^h \partial x^j} = qA^{hj} = \frac{A^{kl}}{2} \left[\frac{\partial B_k}{\partial x^j} H_{lh} + \frac{\partial B_l}{\partial x^h} H_{kh} \right.$$

$$\left. + 2 \left(-\frac{\partial B_l}{\partial x^h} + \frac{1}{2} H_{lh} \right) \left(-\frac{\partial B_k}{\partial x^j} + \frac{1}{2} H_{kj} \right) - 2 \frac{\partial B_l}{\partial x^h} \frac{\partial B_k}{\partial x^j} \right] = \frac{1}{4} A^{kl} H_{lh} H_{kj},$$

the same as that which Mathisson obtains by his method.¹⁶ As he shows, this immediately gives the desired solution: using, this time, the special choice (3.1) of the coefficients A , we only need to take, in (10.1), $h = j = 4$, then $h = j = 1$, thus obtaining

$$\begin{aligned} q &= -H_{14}^2 - H_{24}^2 - H_{34}^2, \\ -q &= H_{41}^2 - H_{21}^2 - H_{31}^2, \end{aligned}$$

and by adding these two equations we see that we must have $H_{24} = H_{34} = H_{21} = H_{31} = 0$; and similarly for the other quantities H . Then, since every H vanishes, $B_h dx^h$ must be an exact differential $-d\mu$ and there exists a transformation (bc) which reduces all the coefficients B to zero, after which C must also be zero, on account of (7.1) and (6.5).

NEW YORK

¹⁶ It is remarkable that Mathisson's deductions and ours follow a quite analogous order, successively obtaining (8.2) and then (10.1), although his parametrix is infinite along a parallel to the x^0 -axis, while our v_0 is a regular function.

A PROOF OF THE FUNDAMENTAL THEOREM ON THE DENSITY OF SUMS OF SETS OF POSITIVE INTEGERS*

BY HENRY B. MANN

(Received January 7, 1942; revised March 16, 1942)

Let $A(B, C)$ be sets of positive integers. Form A^0, B^0 by adjoining 0 to A and B respectively. Let $A(n)$ be the number of positive integers in A that are $\leq n$. The greatest lower bound of the quotients $A(n)/n$ is called the density of A . Let C^0 consist of all integers of the form $a + b(a \in A^0, b \in B^0)$.

Let α be the density of A , β the density of B , γ the density of C , then we shall prove

$$(1) \quad \gamma \geq \alpha + \beta \text{ or } = 1.$$

This inequality has been conjectured by E. Landau, I. Schur and A. Khintchine.

Approximations to (1) have been obtained by the following authors:

E. Landau:¹ $\gamma \geq \alpha + \beta - \alpha\beta$.

A. Besicovitch:² If α^* is the lower bound of the quotients $A(n)/(n + 1)$ then $\gamma \geq \alpha^* + \beta$ or $= 1$.

I. Schur:³ $\gamma \geq \alpha/(1 - \beta)$ or $= 1$.

I. Schur:⁴ $\gamma \geq \frac{1}{2}\alpha + \frac{1}{2}(\alpha^2 + 4\beta^2)^{\frac{1}{2}}$ or $= 1$.

A. Brauer:⁵ $\gamma \geq 9/10(\alpha + \beta)$ or $= 1$.

An important partial result was obtained by A. Khintchine:⁶ If α_i is the density of A_i and $\alpha_1 \leq \alpha_2 \leq \dots \leq \alpha_n$ and $C^0 = \{a_1 + a_2 + \dots + a_n\} (a_i \in A_i^0)$ then $\gamma \geq n\alpha_1$ or $= 1$. In this paper (1) will be proved completely.

Stripped of its transcendental content (1) states that

$$\frac{C(n)}{n} = 1 \text{ or } \geq \min. \frac{A(l)}{l} + \frac{B(m)}{m}, \quad 1 \leq m \leq n, \quad 1 \leq l \leq n.$$

We propose to prove the following sharper theorem.

FUNDAMENTAL THEOREM: Let $A(B, C)$ be sets of positive integers. Let $A(n), B(n), C(n)$ be the numbers of the integers $1, 2, \dots, n$ that are in A, B, C

* Presented to the American Mathematical Society Feb. 28, 1942. The author's interest in this problem was aroused through Dr. A. T. Brauer's lectures on additive theory of numbers at New York University.

¹ *Die Goldbachsche Vermutung und der Schnirelmannsche Satz.* Gottinger Nachrichten, Math. Phys. Klasse (1930), pp. 255-276.

² *On the density of the sum of two sequences of integers.* Jour. London Math. Soc., Vol. 10 (1935), pp. 246-248.

³ *Über den Begriff der Dichte in der additiven Zahlentheorie.* Sitzungsber. der preuss. Akad. der Wiss., Math. Phys. Klasse, (1936), pp. 269-297.

⁴ L.c. footnote 3.

⁵ These Annals, Vol. 42, (1941), pp. 959-988.

⁶ *Zur additiven Zahlentheorie.* Matematičeski Sbornik, Vol. 39, (1932) pp. 27-34.

respectively. Let $C^0 = \{a + b\} (a \in A^0, b \in B^0)$ where A^0, B^0 consist of the numbers of A and B respectively and the number 0. Then

$$(2) \quad \frac{C(n)}{n} = 1 \quad \text{or} \quad \geq \min. \frac{A(m) + B(m)}{m}, \quad 1 \leq m \leq n, \quad m \notin C.$$

Let $n_1 < n_2 \dots$ be the numbers of \bar{C} (i.e. numbers not in C); then $C(n_r) = n_r - r$. (2) is an immediate consequence of the inequality

$$(3) \quad \frac{C(n_r)}{n_r} \geq \min. \frac{A(n_q) + B(n_q)}{n_q}, \quad q \leq r.$$

We divide the numbers $\leq n_r$ into three sets: numbers of B , numbers of the form $n_r - a (a \in A^0)$, numbers of L_r where L_r denotes the set of all positive numbers $\leq n_r$ that are neither in B nor of the form $n_r - a (a \in A^0)$. Denote by l_r the number of integers in L_r . These three sets are disjoint. Hence

$$(4) \quad n_r = B(n_r) + A(n_r) + 1 + l_r.$$

(3) follows from (4) for $r = 1$. In proving (3) by induction we may assume

$$\frac{C(n_s)}{n_s} > \frac{C(n_r)}{n_r} \quad \text{or} \quad rn_s > sn_r \quad \text{for} \quad s = 1, \dots, r-1.$$

Our statement then is that

$$(5) \quad \frac{A(n_s) + B(n_s)}{n_s} > \frac{C(n_r)}{n_r} \quad \text{for} \quad s = 1, \dots, r-1$$

implies

$$(6) \quad C(n_r) \geq A(n_r) + B(n_r),$$

or using (4)

$$\frac{1 + l_s}{n_s} < \frac{r}{n_r} \quad \text{for} \quad s = 1 \dots r-1 \quad \text{implies} \quad 1 + l_r \geq r.$$

We have to prove the following lemma: If $rn_s > sn_r, rn_s > (1 + l_s)n_r$ for $s = 1, 2, \dots, r-1$, then $l_r \geq r-1$.

From now on r is fixed. We therefore write $n_r = N, L_r = L, l_r = l$. s, t range from 1 to $r-1$. Put $d_s = N - n_s$ then $n_s - d_s = n_t - d_t$.

Construction of numbers of L . To prove the lemma we have to construct $r-1$ numbers b' in L , i.e. numbers not in B and not of the form $N - a (a \in A^0)$. To satisfy both conditions we construct numbers b' satisfying the equations

$$b' = n_s - a = N - \bar{a} \quad (a \in A^0, \bar{a} \notin A^0).$$

The condition \bar{a} not $\in A^0$ is satisfied by putting $\bar{a} = n_t - b$ or

$$b' = b + d_t = n_s - a.$$

Thus we arrive at the following recursive definition of numbers e_1, e_2, \dots and sets B^0, B^1, B^2, \dots , and T_1, T_2, \dots , starting with B^0 .

DEFINITION: a) We choose e_{i+1} as an element in $B^0 \cup B^1 \cup \dots \cup B^i$.

b) t lies in T_{i+1} if there is an $a \in A^0$ such that

$$(7) \quad a + e_{i+1} + d_i = n_i$$

and neither s nor t in $T_1 \cup T_2 \cup \dots \cup T_i$.

c) $B^{i+1} = \{e_{i+1} + d_i\}_{t \in T_{i+1}}$.

Equation (7) is equivalent to

$$(7') \quad a + e_{i+1} + d_i = n_i,$$

hence s as well as t lies in T_{i+1} . We may therefore write

$$(8) \quad b' = e_{i+1} + d_i = n_i - a \in B^{i+1}.$$

The following propositions follow easily from the definition:

1. $e_j \in B^0 \cup B^1 \cup \dots \cup B^{j-1}$ ($j \geq 1$).
2. No element of T_j lies in $T_1 \cup T_2 \cup \dots \cup T_{j-1}$.
3. $B^j = \{e_j + d_i\}_{t \in T_j}$.

COROLLARY: B^j contains as many elements as T_j . All elements of B^j are greater than e_j .

4. An element b' of B^j may be written as $n_s - a$ ($s \in T_j, a \in A^0$). Also n_s can be decomposed into $a + b'$ ($a \in A^0, b' \in B^j$).

5. If $a + b' = n_s$, ($a \in A^0, b' \in B^0 \cup B^1 \cup \dots \cup B^j$) then $s \in T_1 \cup \dots \cup T_j$.

PROOF: Proposition 5 is true for $j = 0$. We assume that it is true for $j = i$ and prove it for $j = i + 1$. If $b' \in B^{i+1}$ then $b' = e_{i+1} + d_i$, $t \in T_{i+1}$, hence $t \notin T_1 \cup T_2 \cup \dots \cup T_i$, therefore $a + e_{i+1} + d_i = n_s$, $a + e_{i+1} + d_i = n_i$. If $s \notin T_1 \cup T_2 \cup \dots \cup T_i$, then $s \in T_{i+1}$ according to definition of T_{i+1} .

6. No element of B^j ($j \geq 1$) lies in $B^0 \cup B^1 \cup \dots \cup B^{j-1}$ nor is of the form $N - a$ ($a \in A^0$).

PROOF: An element b_j of B^j is of the form $n_s - a$ ($s \in T_j, a \in A^0$). But the relation

$$n_s = a + b_j, \quad s \notin T_1 \cup T_2 \cup \dots \cup T_{j-1}, \quad b_j \in B^0 \cup B^1 \cup \dots \cup B^{j-1}$$

contradicts 5. Each element b_j of B^j is of the form $e_j + d_s$ ($s \in T_j$): the equation $e_j + d_s = N - a$, ($a \in A^0$) or $a + e_j = n_s$ contradicts 5 because of

$$s \notin T_1 \cup T_2 \cup \dots \cup T_{j-1}, \quad e_j \in B^0 \cup B^1 \cup \dots \cup B^{j-1}.$$

In order to arrive at a uniquely determined construction we shall choose e_{i+1} as the least number for which T_{i+1} is not empty. The final construction is then defined as follows:

a) t lies in $T_{i+1}\{e\}$ if there is an $a \in A^0$ such that

$$a + e + d_i = n_i$$

and neither t nor s in $T_1 \cup \dots \cup T_i$.

b) e_{i+1} is the smallest e in $B^0 \cup B^1 \cup \dots \cup B^i$ for which $T_{i+1}\{e\}$ is not empty.

c) $T_{i+1} = T_{i+1}\{e_{i+1}\}$, $B^{i+1} = \{e_{i+1} + d_i\}_{t \in T_{i+1}}$.

The process is to be continued until $T_{j+1}\{e\}$ is empty for every $e \in B^0 \cup B^1 \cup \dots \cup B^j$.

We are then able to prove the following propositions:

7. $T_{j+1}\{e\}$ is empty for every $e \in B^0 \cup B^1 \cup \dots \cup B^j$.
8. If $e \in B^0 \cup B^1 \cup \dots \cup B^{j-1}$ and $T_j\{e\}$ is not vacuous then $e \geq e_j$.
9. $e_1 < e_2 < \dots < e_j$.

PROOF: $e_j = e_{j+1}$ is impossible because $T_{j+1}\{e_j\}$ is empty. $e_j < e_{j+1}$ means therefore that there is no $e < e_j$ for which $e \in B^0 \cup \dots \cup B^j$ and $T_{j+1}\{e\}$ is not empty. If there were, then $T_j\{e\}$ would not be empty. Either $e \in B^0 \cup \dots \cup B^{j-1}$, and then $e \geq e_j$ according to 8; or $e \in B^j$, and then $e > e_j$ according to the corollary to 3.

10. All elements of $B^j \cup \dots \cup B^J$ are greater than e_j ($1 \leq j \leq J$). This follows from the corollary to 3 and from 9.

11. Let s be the least index not in $T_1 \cup T_2 \cup \dots \cup T_J$. Let there be q indices $(r-1, r-2, \dots, r-q)$ such that $d_i \leq n_s$. Then

$$l_s \geq q.$$

PROOF: $n_s - d_i$ ($d_i \leq n_s$) is either in C^0 and then $= a + b'$ ($a \in A^0$, $b' \in B^0$) or $= n_u$, ($u \in T_1 \cup \dots \cup T_J$) and hence $= a + b'$ ($a \in A^0$, $b' \in B^0 \cup \dots \cup B^J$). Thus in either case

$$n_s - d_i = a + b' \quad (a \in A^0, b' \in B^0 \cup \dots \cup B^J).$$

If t is not in $T_1 \cup \dots \cup T_J$ then $T_{J+1}\{b'\}$ is not empty which contradicts 7.

Let $t \in T_j$, $e_j + d_i = b_j \in B^j$. We assert $e_j + d_i \leq n_s$. If $e_j + d_i > n_s$, then

$$e_j > n_s - d_i = a + b' \geq b'.$$

Thus b' cannot lie in $B^j \cup \dots \cup B^J$ (see Prop. 10). Hence $b' \in B^0 \cup \dots \cup B^{j-1}$. The equation

$$a + b' + d_i = n_s$$

with $t \notin T_1 \cup \dots \cup T_{j-1}$, $s \notin T_1 \cup \dots \cup T_{j-1}$ shows that $T_j\{b'\}$ is not vacuous and $b' < e_j$ is impossible by Prop. 8.

b_j is not in B nor of the form $n_s - a$ ($a \in A^0$) (see 5). Hence it is in L_s . Because of 6 we obtain q distinct such numbers b_j corresponding to the q values of t . Therefore

$$l_s \geq q.$$

We can prove now our lemma. Suppose that

$$rn_i > iN, \quad rn_i > (1 + l_i)N \quad \text{for } i = 1, 2, \dots, r-1.$$

We shall show that $T_1 \cup \dots \cup T_J$ contain all indices $i \leq r-1$. Because of 6 and the corollary to 3 our lemma is an immediate consequence of this fact. Suppose s is the least index not in $T_1 \cup \dots \cup T_J$. Let q be defined as in Proposition 11. Then

$$rn_s > (1 + l_s)N \geq (q + 1)N, \quad rn_{(r-q-1)} > (r - q - 1)N.$$

Adding these two inequalities we obtain

$$n_s + n_{(r-q-1)} > N \quad \text{or} \quad n_s > d_{(r-q-1)}.$$

But then we would have $q + 1$ indices $t = (r - 1, r - 2, \dots, r - q - 1)$ for which $d_t \leq n$, contrary to the significance of q . This proves our lemma.

Another result of a different nature can be obtained by the construction method used in the proof of the fundamental theorem. We propose to prove

THEOREM II. *If $C^0 = \{a + b\} (a \in A^0, b \in B^0)$ and if α^* is the G.L.B. of the quotients $A(m)/(m + 1)$ and if n is not in C . Then*

$$(9) \quad C(n) \geq \alpha^* n + B(n).$$

From equation (4) we have for $r = 1$.

$$C(n_1) \geq \alpha^* n_1 + B(n_1) + (\alpha - \alpha^*) n_1.$$

We propose to prove by induction

$$(9') \quad C(n_i) \geq \alpha^* n_i + B(n_i) + (\alpha - \alpha^*) n_1, \quad i = 1, 2, \dots$$

We distinguish two cases: 1). $n_r - n_{(r-1)} > 1$; 2). $n_r - n_{(r-1)} = 1$.

PROOF FOR CASE 1: The integers $> n_{(r-1)}$ and $\leq n_r$ are either in B or of the form $n_r - a$ ($a \in A^0, 0 \leq a < n_r - n_{(r-1)}$) or in neither of these two sets. Hence

$$\begin{aligned} n_r - n_{(r-1)} &\geq A(n_r - n_{(r-1)} - 1) + B(n_r) - B(n_{(r-1)}) + 1 \\ &\geq \alpha^* (n_r - n_{(r-1)}) + B(n_r) - B(n_{(r-1)}) + 1. \end{aligned}$$

By induction we have

$$C(n_{(r-1)}) = n_{(r-1)} - (r - 1) \geq \alpha^* n_{(r-1)} + B(n_{(r-1)}) + (\alpha - \alpha^*) n_1.$$

Adding both inequalities we obtain (9') for $i = r$.

PROOF FOR CASE 2: In this case we have $d_{(r-1)} = 1, n_1 - d_{(r-1)} = a + b$ ($a \in A^0, b \in B^0$). We form $T_1\{b\}$ and B^1 . Let B^* contain the numbers of B^0 and B^1 . Put $C^* = \{a + b'\} (a \in A^0, b' \in B^*)$. n_r will then be the j^{th} gap $j < r$ in C^* and we have by induction

$$C^*(n_r) \geq \alpha^* n_r + B^*(n_r) + (\alpha - \alpha^*) n_1.$$

But by proposition 5 we have

$$C^*(n_r) - C(n_r) = B^*(n_r) - B(n_r).$$

Subtracting this equation we obtain (9') for $i = r$.

It is not possible to substitute α for α^* in (9) as shown by the following example

$$\begin{aligned} A = B &= \{1, 2, 6, 7, 8, 12, \dots\}, \\ C &= \{1, 2, 3, 4, 6, 7, 8, 9, 10, 12, \dots\}, \\ \alpha &= \frac{2}{3}, \quad B(11) = 5, \quad C(11) = 9. \end{aligned}$$

If $C^0 = \{a + b\} (a \in A^0, b \in B)$ and if $1 \in B$ then it can be shown exactly by the same method that $C(n) \geq \alpha^* n + B(n)$ if n is not in C .

A REMARK ON S. KAKUTANI'S CHARACTERIZATION OF (L) -SPACES¹

By M. F. SMILEY

(Received March 9, 1942)

The purpose of this note is to show that we may rephrase the condition (IX) of Kakutani in a way which indicates a close relation between the norm in an (AL) -space and the metric of a metric lattice.² This will also permit a slight weakening of assumptions in §§3-5.

We show first that the condition

$$(IX') \quad \|x - y\| = \|x \vee y - x \wedge y\|$$

is equivalent to (IX) under (I)-(IV), (VI)-(VII).

PROOF. (IX) *implies* (IX'): By Lemma 1.5, $x = x_+ - x_-$ and $x_+ \wedge x_- = 0$; consequently $\|x\| = \|x_+ - x_-\| = \|x_+ + x_-\|$ by (IX). Since $-x \vee 0 = -(x \wedge 0)$, we have the equation

$$\|x\| = \|x \vee 0 + (-x \vee 0)\| = \|x \vee 0 - x \wedge 0\|.$$

Replacing x by $x - y$ and applying Lemma 1.2 yields (IX').

(IX') *implies* (IX): If $x \wedge y = 0$, then $x \vee y = x + y$ by Lemma 1.3. Hence, if $x \wedge y = 0$, the condition (IX') yields $\|x - y\| = \|x \vee y\| = \|x + y\|$.

We may now prove that (I)-(IV), (VI)-(IX) imply the following condition.

$$(1) \quad \|a \vee x - a \vee y\| + \|a \wedge x - a \wedge y\| = \|x - y\|.$$

PROOF.³ By (IX') and Lemma 1.6 the left side of (1) is $\|a \vee x \vee y - a \vee (x \wedge y)\| + \|a \wedge (x \vee y) - a \wedge x \wedge y\|$. Condition (VIII) reduces this to the expression

$$\|a \vee x \vee y - a \vee (x \wedge y) + a \wedge (x \vee y) - a \wedge x \wedge y\|.$$

Using Lemma 1.3 we obtain the identity (1) at once.

The identity (1) shows that the assumption (V) is not needed in §§3-5. We may even show that the triangle inequality for the norm is a consequence of (I)-(IV), (VI)-(IX) by a simple paraphrase of the proof of Theorem 3.10 of G. Birkhoff.⁴

¹ Presented to the American Mathematical Society, April 3, 1942. Our title refers to Kakutani's paper *Concrete representation of abstract (L) -spaces and the mean ergodic theorem*, these Annals, vol. 42 (1941), pp. 523-537. All references are to this paper unless the contrary is explicitly stated.

² Cf. G. Birkhoff, *Lattice Theory*, Amer. Math. Soc. Col. Pub., vol. 25 (1940), p. 41.

³ *Ibid.*, p. 42.

⁴ *Ibid.*, p. 42.

NOTE ADDED IN PROOF. G. Birkhoff has remarked to the author that $m[a] = \|a_+\| - \|a_-\|$ is a sharply positive modular functional in an (L)-space. This result depends only on conditions (I)–(IV), (VI)–(VIII). For, if $a \geq b$, then $m[a] - m[b] = \|a - b\|$ by (VIII) and Lemma 1.2. Since $\|a\| > 0$ unless $a = 0$, this shows that $m[a]$ is sharply positive. Applying this result again we find that $m[a] - m[a \wedge b] = \|a - a \wedge b\| = \|a \vee b - b\| = m[a \vee b] - m[b]$. Thus $m[a]$ is also a modular functional. Our condition (IX') then merely requires that the metric $\delta(a, b)$ which arises from $m[a]$ should coincide with $\|a - b\|$. When this is true, the identity (1) is a consequence of (in fact, equivalent to) the distributive law.

LEHIGH UNIVERSITY.

GENERALIZED SURFACES IN THE CALCULUS OF VARIATIONS. II

By L. C. YOUNG

(Received December 21, 1941)

MEAN SURFACES AND THE THEORY OF THE PROBLEM $\iint f(x, y, p, q) dx dy = \text{Min.}$

1. Introduction

In this second note, machinery is produced for studying extensions of the classical necessary conditions applicable both in the ordinary and in the generalized minimum problem for surface integrals. We employ it to derive necessary and sufficient conditions in the case of an integrand $f(x, y, p, q)$ in which z does not occur explicitly. Our methods follow naturally from the idea of generalized surface developed in the first note,¹ but are no longer restricted to continuous integrands $f(x, y, z, p, q)$. We show, in fact, that for a much wider class of these integrands the minimum in the problem considered is unaltered if we admit, in addition to ordinary surfaces, only a special kind of generalized surface which we term "mean surface." The ideas connected with the notion of a mean surface are closely akin to a method developed by Haar² in the special case of a regular problem. They lead to a generalization of an inequality, due to Steiner,³ valid when the integrand has the form $f(x, y, p, q)$; and thence to equations characterizing a Lipschitzian surface for which our minimum is attained.

2. Mean surfaces

Let z_1 and z_2 denote the tracks, and $M_1(g)$ and $M_2(g)$ the averages at x, y , of two generalized surfaces S_1 and S_2 or, in particular, of two ordinary surfaces. We define the *mean* of S_1 and S_2 as the generalized surface S with the track $z = \frac{1}{2}z_1 + \frac{1}{2}z_2$, and with the average $M(g) = \frac{1}{2}M_1(g) + \frac{1}{2}M_2(g)$ at the point x, y . A generalized surface which is the mean of two *ordinary* surfaces will be termed shortly a *mean surface*.

(2.1) *Any mean surface is expressible as the mean of two ordinary surfaces whose tracks differ by at most ϵ .*

To prove this, it is clearly sufficient to show that if S is the mean of two ordinary surfaces S_1 and S_2 whose tracks z_1 and z_2 differ by at most $2a$, then S is also expressible as the mean of a second pair of ordinary surfaces whose tracks differ by at most a . Let E, E', E'' denote the sets of x, y in which, respectively,

$$|z_1 - z_2| < a, \quad z_1 - z_2 \leq -a, \quad z_1 - z_2 \geq a;$$

¹ Young [12].

² Haar [3].

³ Steiner [10], p. 298.

and let \bar{z}_1, \bar{z}_2 denote the pair of functions which coincide with

$$z_1, z_2 \text{ in } E, \quad z_2 - a, z_1 + a \text{ in } E', \quad z_2 + a, z_1 - a \text{ in } E''.$$

We verify at once that these functions coincide with z_1, z_2 at boundary points of E , since we then have $z_2 - z_1 = \pm a$; it follows easily that they satisfy a Lipschitz condition with a same constant as the functions z_1 and z_2 : for any segment can be decomposed into at most three, each of which has both ends belonging to the closure of the same set E, E' , or E'' , and in these partial segments the Lipschitz condition for the new functions is identical with the original one for z_1, z_2 or their permutation z_2, z_1 .

Finally it is clear that the new functions have the same gradient as the original ones or their permutation, wherever the four gradients exist, and so almost everywhere; and since we have also

$$\frac{1}{2}\bar{z}_1 + \frac{1}{2}\bar{z}_2 = \frac{1}{2}z_1 + \frac{1}{2}z_2 \quad \text{and} \quad |\bar{z}_2 - \bar{z}_1| \leq a,$$

this completes the proof.

3. Integrals over mean surfaces. Comparison of minima

We shall be dealing with surface integrals of functions $f(x, y, z, p, q)$ measurable B but not necessarily continuous. The average $M(g)$ at x, y which constitutes part of the definition of generalized surface, must now be suitably extended to discontinuous functions $g(p, q)$. In the cases that we shall be concerned with, this extension is either trivial, or else sufficiently treated by well-known classical writers.

As in the first note, all surfaces which occur will be supposed Lipschitzian.

In the case of a mean surface S , the average $M(g)$ is of the form $\frac{1}{2}g(p', q') + \frac{1}{2}g(p'', q'')$, where p', q' and p'', q'' denote positions of p, q dependent on x, y , which will be termed the two *possible positions* of p, q at x, y . In accordance with our definition of mean surface, these possible positions are further restricted to be at almost every x, y of the domain A a permutation of the gradients of at least one fixed pair of ordinary surfaces. The integral $F(S)$ of $f(x, y, z, p, q)$ over the mean surface S is obtained by writing

$$\iint_S f(x, y, z, p, q) dx dy = \iint_A \frac{1}{2} \{f(x, y, z, p', q') + f(x, y, z, p'', q'')\} dx dy,$$

where on the right z denotes the track of S .

The notions of ordinary surface, mean surface, generalized surface give rise to three different problems of minimum, obtained by interpreting correspondingly the class of admissible surfaces with a given boundary for which we seek to make $F(S)$ a minimum. If the corresponding minimal values of $F(S)$ are denoted by m_o, m_m, m_g , it is evident that

$$(3.1) \quad m_o \geq m_m \geq m_g,$$

since every ordinary surface is a mean surface (the mean of two coincident, or parallel, ordinary surfaces), while every mean surface is a generalized surface.

The results of the first note⁴ show that the inequalities (3.1) reduce to equalities when f is continuous, since the extreme terms then have the same value. On the other hand, it is easily seen that at least one of the inequalities is strict for certain discontinuous functions f : thus if we write $f = -1$ when p, q equals either $-y, x$ or $y, -x$, and $f = 0$ otherwise, we obtain a function $f(x, y, p, q)$ which vanishes almost everywhere on any ordinary surface, but which has the average $M(f) = -1$ at all points of the generalized surface with the track $z = 0$ and the average $M(g)$ at x, y given by the expression $\frac{1}{2}g(-y, x) + \frac{1}{2}g(y, -x)$; so that m_o vanishes while m_o is negative in this case.

We shall show, in the next paragraph, that for a class of functions including among others all those of the form $f(x, y, p, q)$ which are bounded in bounded sets, and in particular the function just considered, the first of the inequalities (3.1) is necessarily an equality, i.e. we have

$$(3.2) \quad m_o = m_m.$$

4. Equality of the minima m_o and m_m

We now come to the first main theorem of this note, which will enable us to treat the ordinary problem on the same footing as the very much simpler generalized problem from the point of view of necessary conditions. This theorem asserts that (3.2) is true under general conditions. Since this equality has already been established for continuous integrands $f(x, y, z, p, q)$, the reader whose interest is limited to classical problems may feel inclined to omit its proof. It should be noted, however, that it is frequently desirable to transform a classical integrand into a discontinuous one by moving the origin of z, p, q to a position which depends on x, y . Such a transformation can be used, for instance, to render certain results of the present note applicable, not only to Lipschitzian, but also to more general surfaces. If we do not insist on these extensions, it is on account of the greater simplicity of the statements for the Lipschitzian case and because we have no evidence that conditions derived from these by means of the above transformation would be essential ones. It is also, perhaps, interesting and instructive to see how the necessary conditions of Euler and Weierstrass, once they have been freed of the formal differential apparatus, can be applied to integrands which are not even continuous.

(4.1) *Suppose firstly that $f(x, y, z, p, q)$ is a function measurable B which is upper-semicontinuous in z when the other variables are fixed;⁵ and secondly that this function is bounded above⁶ for any relevant bounded system of x, y, z, p, q . Then, given any mean surface S and real numbers k and ϵ , where $k > F(S)$ and $\epsilon > 0$, there is an ordinary surface S' with the same boundary as S , such that*

⁴ Young [12], (14.1), p. 102.

⁵ This is in particular the case when f is independent of z .

⁶ This second condition may be relaxed: it is enough to suppose that in A the function f is less than an integrable function of x, y only. Moreover, when f is independent of z , we require this only in the neighbourhood of the boundary of A .

$$(4.2) \quad F(S') < k,$$

and such that, moreover, its track differs from that of S by at most ϵ , and its Lipschitz constant exceeds that of S by at most ϵ .

Denoting by n an arbitrary positive integer and by E the set of the points x, y of A whose distance from the boundary of A is not less than $1/n$, we can, by (2.1) and McShane's lemma,⁷ construct two ordinary surfaces S' and S'' having S for their mean in the set E , in such a manner that the tracks of S' and S'' differ by at most $1/n^2$ in E and coincide with that of S on the boundary of A , while at the same time their Lipschitz constants exceed by at most $1/n$ that of S . We shall suppose that $F(S') \leq F(S'')$.

Now the tracks z' and z'' of our two surfaces tend to the track z of S as n tends to infinity. Moreover, any point x, y of A becomes a point of E if n is sufficiently large, provided that x, y is not a boundary point; therefore at almost every point of A the gradients p', q' and p'', q'' of our two surfaces are a permutation of the possible positions of p, q for the mean surface S , provided that n is large enough. Hence the average $M(f)$ for S is almost nowhere less than the upper limit as n tends to infinity of the expression

$$\frac{1}{2}\{f(x, y, z', p', q') + f(x, y, z'', p'', q'')\},$$

by semicontinuity of f . It follows from a suitable form of Fatou's theorem⁸ that

$$\iint_A M(f) dx dy \geq \liminf_n \iint_A \frac{1}{2}\{f(x, y, z', p', q') + f(x, y, z'', p'', q'')\} dx dy$$

and so, that $k > \frac{1}{2}F(S') + \frac{1}{2}F(S'')$, if n is sufficiently large; clearly this requires $k > F(S')$. The remaining assertions are satisfied if we choose a value of n exceeding ϵ^{-2} . This completes the proof.

5. A convexity property of the minimum

From now on, it will be assumed that the integrand of the problem is a function $f(x, y, p, q)$ which is independent of the variable z . This function will be supposed moreover measurable B and, for the present, bounded above when x, y lies in A and p, q in any bounded set.

Denoting by U a function defined on the boundary of A in such a manner that there exists a Lipschitzian surface whose boundary is given by U , we introduce the functionals $\varphi_0(U)$, $\varphi_0(U)$, $\varphi_0(U, K)$ and $\varphi_0(U, K)$, any one of which we designate shortly $\varphi(U)$. We define these to be absolute minima of our surface-integral for the surfaces with boundary U which belong respectively to the following four classes:

- ordinary Lipschitzian surfaces;
- generalized Lipschitzian surfaces;
- ordinary surfaces with Lipschitz constants less than K ;
- generalized surfaces with Lipschitz constants less than K .

⁷ Young [12], (9.1), p. 93.

⁸ Saks [8], (12.10) p. 29, in the form obtained by taking a sequence $a - f_n$ instead of a non-negative sequence f_n .

(5.1) *Each of the minima $\varphi(U)$ is convex in U .*

Before proving this, let us state for completeness, that we term *convex* a functional $Q(U)$ in an arbitrary functional space, if given any two points U' and U'' of this space at which this functional is defined and finite above, and given further any real number θ where $0 \leq \theta \leq 1$, the functional is defined at the point $\theta U' + (1 - \theta)U''$ and satisfies

$$(5.2) \quad Q[\theta U' + (1 - \theta)U''] \leq \theta Q(U') + (1 - \theta)Q(U'').$$

It is known that this definition is equivalent to the following conditions:

- the left-hand side of (5.2) is a function of θ bounded above on the unit segment;
- the inequality (5.2) holds for $\theta = \frac{1}{2}$.⁹

We shall use this equivalent form of the definition in proving (5.1).

We observe in the first place that if U' and U'' are boundaries of Lipschitzian surfaces belonging to the relevant class, then the same is the case of $\theta U' + (1 - \theta)U''$ when θ lies on the unit segment *and indeed when θ lies on a segment including the unit segment in its interior*. The left-hand side of (5.2), when we choose $Q(U) = \varphi(U)$, is majorized by the values of $F(S)$ for certain surfaces with bounded Lipschitz constants, and so is certainly bounded above as a function of θ on the segment in question.

It remains to verify the second condition. Let k be any number exceeding $\frac{1}{2}\varphi(U') + \frac{1}{2}\varphi(U'')$ and therefore exceeding $\frac{1}{2}F(S') + \frac{1}{2}F(S'')$ where S' are certain suitably chosen admissible surfaces with boundaries U' and U'' respectively. The number k then exceeds $F(S)$, where S denotes the mean of S' and S'' . But S has the boundary $\frac{1}{2}U' + \frac{1}{2}U''$ and S is either an admissible surface, or can be added to the admissible surfaces without altering our minimum, and so $\varphi(\frac{1}{2}U' + \frac{1}{2}U'') \leq F(S) < k$. This completes the proof of (5.1).

In the case of the generalized problem, the theorem just proved has a variant which reduces in the *regular* case to a well-known inequality of Haar-Steiner.¹⁰ We denote by $I(z)$ the minimum of $F(S)$ for surfaces S with the track z , and we note that in the regular case $I(z)$ reduces to the value of $F(S)$ for the ordinary surface with this track.

(5.3) *The functional $I(z)$ is convex.*¹¹

To prove this, we select an arbitrary real number θ on the unit segment and two Lipschitzian tracks z' and z'' . We denote by S' and S'' two admissible generalized surfaces with these tracks, and by $M'(g)$, $M''(g)$ the corresponding averages at x , y ; and we write S for the generalized surface with the track $\theta z' + (1 - \theta)z''$ and the average $\theta M' + (1 - \theta)M''$ at x , y . Then

$$I(\theta z' + [1 - \theta]z'') \leq F(S) = \theta F(S') + [1 - \theta]F(S''),$$

⁹ The equivalence of the definitions easily reduces to the case of real functions treated by Hardy, Littlewood and Pólya [6].

¹⁰ Haar [3], p. 227, Inequality (3).

¹¹ We assert this only for the generalized problem. The same is then evidently true of the ordinary problem *only in the regular case*.

from which it follows, passing to the lower bound for all S' and S'' , that

$$I(\theta z' + [1 - \theta]z'') \leq \theta I(z') + [1 - \theta]I(z'').$$

This completes the proof.

For our purposes, a second variant of (5.1), valid this time for ordinary admissible surfaces as well as for generalized ones, is required. This variant constitutes a minor extension of the original theorem.

6. Prescribed discontinuities¹²

The classical minimum problems for continuous surfaces may be regarded as special cases of corresponding problems for more general surfaces whose discontinuities are specified. We shall require the extension of (5.1) to a problem of this type, in which we admit piecewise Lipschitzian surfaces with prescribed discontinuities along given parallels to the coordinate axes.

By a *piecewise Lipschitzian surface* S (ordinary or generalized) we shall mean the aggregate of a finite number of Lipschitzian portions of surface S_{ik} (ordinary or generalized) defined on partial domains A_{ik} of a subdivision of A by a finite number of parallels to the axes. It will be understood that S is regarded as unaltered by subsequent subdivision of its constituent Lipschitzian portions. By the *surface integral* $F(S)$, we shall mean similarly the sum of the surface integrals $F(S_{ik})$ over these portions, and this definition is again invariant under further subdivision.

By the *discontinuity of* S , we shall mean a function $V(x, y)$ only defined at interior points of A and possibly undefined at a finite number of these, such that V vanishes at interior points of each A_{ik} while on the line separating two adjacent partial domains $A_{i+1,k}$ and A_{ik} or else $A_{i,k+1}$ and A_{ik} , the function V is the difference of the tracks of the corresponding Lipschitzian portions. Besides this discontinuity V , we have a boundary-function U just as in the case of a continuous surface.

It is convenient to combine the two functions U on the boundary of A and V in the interior of A , into a single function W defined throughout A (except possibly at a finite number of points). The function W then specifies the boundary and discontinuity of S .

(6.1) *The functions W which correspond in this way to at least one such piecewise Lipschitzian surface S constitute a vector-space.*

For if S' and S'' are two piecewise Lipschitzian surfaces and W' and W'' specify their boundaries and discontinuities, we can divide A into portions A_{ik} corresponding to continuous portions of both surfaces simultaneously. Denoting by z'_{ik} and z''_{ik} the tracks of these portions we define a piecewise Lipschitzian surface S consisting of the continuous portions of ordinary surfaces

¹² We do not study this problem for its own sake, but in order to obtain machinery for the special case in which there are no discontinuities. In the literature only problems with *unspecified* discontinuities appear to have been treated.

given by the track $az'_{ik} + bz''_{ik}$ on A_{ik} . This surface S then has the boundary and discontinuity specified by the function $W = aW' + bW''$, and this proves (6.1). Moreover, we see that S has in each of its continuous portions a Lipschitz constant not exceeding $|a| + |b|$ times the greatest of those of S' and S'' , so that we have the following result:

(6.2) *If W' and W'' correspond to piecewise Lipschitzian surfaces whose continuous portions have Lipschitz constants less than K , then $aW' + bW''$ corresponds to at least one piecewise Lipschitzian surface whose continuous portions have Lipschitz constants less than $(|a| + |b|)K$.*

This being so, we define by analogy with §5 the functionals of W consisting of the absolute minima of $F(S)$ for the piecewise Lipschitzian surfaces S with boundary and discontinuity specified by W which can be subdivided into continuous portions belonging to the same four classes as previously. As these functionals reduce to those of §5 when the variable W reduces to U , i.e. when there is no discontinuity, we shall designate them by the symbols $\varphi_o(W)$, $\varphi_e(W)$, $\varphi_o(W, K)$ and $\varphi_e(W, K)$, and any one by the symbol $\varphi(W)$.

As an immediate consequence of (6.2) we have:

(6.3) *Any two points of the W -space for which $\varphi(W) \neq +\infty$ are inner points of a segment on which $\varphi(W)$ is bounded above.*

Moreover, by exactly the same argument as that used to prove the corresponding inequality in §5, we see that

$$(6.4) \quad \varphi(\tfrac{1}{2}W' + \tfrac{1}{2}W'') \leq \tfrac{1}{2}\varphi(W') + \tfrac{1}{2}\varphi(W'').$$

Combining (6.3) and (6.4), we deduce:

(6.5) *Each of the minima $\varphi(W)$ is convex in W .*

7. Existence of a "flat support"

The property of convexity of the functionals $\varphi(U)$, $I(z)$, $\varphi(W)$ owes its importance to the fact that it places at our disposal the powerful theorem of Minkowski asserting the existence of a "flat support." We shall follow Banach¹³ in proving a form of this theorem valid in general vector spaces which covers our needs.

We consider a space whose elements are abstract vectors X . As is customary, a functional $L(X)$ is termed linear if $L(cX) = cL(X)$ for any constant c , and $L(X + Y) = L(X) + L(Y)$; and since this definition implies that $L(0) = 0$, it is frequently necessary to introduce an additive constant when the analogy with Euclidean space is to be preserved.

The expression $a + L(X)$, where a is a constant and $L(X)$ a linear functional, is termed "*flat support*" at X_0 of a given functional $Q(X)$, if it never exceeds $Q(X)$ and if it takes the value $Q(X_0)$ at X_0 .

¹³ Banach [1], p. 27, Theorem 1.

One more definition: if G is a vector subspace, we denote by $\{X_0, G\}$ the vector space or subspace which consists of all vectors Y of the form $X + tX_0$ where X lies in G and t is real. It is easy to see that if X_0 does not lie in G , this expression of a vector Y of $\{X_0, G\}$ is *unique*. We shall denote further by $\{X_0, G\}_+$ and $\{X_0, G\}_-$ the systems of vectors Y of the above form with t positive, and with t negative, respectively.

(7.1) *Let $Q(X)$ be a convex functional defined in a vector-space, and let $L(X)$ be a linear functional defined in a vector subspace G and such that throughout G we have $L(X) \leq Q(X)$; suppose further that X_0 is any point of the vector-space such that $Q(X)$ is defined and finite above for at least one X of each of the sets $\{X_0, G\}_+$ and $\{X_0, G\}_-$. Then we can extend the definition of $L(X)$ in such a manner that this functional be defined, linear, and not greater than $Q(X)$ throughout $\{X_0, G\}$.*

In both the hypothesis and the conclusion of the theorem, the inequality connecting $L(X)$ and $Q(X)$ is regarded as automatically true at points X for which $Q(X)$ is either not defined or not finite above.

In proving (7.1), we may clearly suppose that X_0 does not lie in G . Denoting by k_0 a number to be fixed later, we continue $L(X)$ by writing, for real t and for arbitrary X in G ,

$$L(X + tX_0) = L(X) + tk_0,$$

and this continuation is linear in $\{X_0, G\}$, so that we have only to show that it does not exceed $Q(X + tX_0)$. Since this last condition holds for $t = 0$, we need only verify it for $t = a$ and $t = -a'$, where a, a' are positive; it becomes

$$(7.2) \quad \{L(X) - Q(X - a'X_0)\}/a' \leq k_0 \leq \{Q(X' + aX_0) - L(X')\}/a$$

where X and X' are arbitrary vectors of G .

It only remains to establish the existence of a finite k_0 which satisfies (7.2) for all a, a', X, X' . Moreover, since Q is defined and finite above *somewhere* in each of the sets $\{X_0, G\}_+$ and $\{X_0, G\}_-$, we see that the upper bound of the expression on the extreme left of (7.2) is not $-\infty$, and the lower bound of the expression on the extreme right is not $+\infty$. The existence of k_0 is thus ensured if we have an inequality between these bounds, which is equivalent to the statement that every value of the extreme left of (7.2) is less than or equal to every value on the extreme right.

Now this last statement, after multiplying by $aa'/(a + a')$, reduces to the inequality

$$\frac{aL(X) + a'L(X')}{a + a'} \leq \frac{aQ(X - a'X_0) + a'Q(X' + aX_0)}{a + a'}$$

and the latter is satisfied since, if X'' is the vector $(aX + a'X')/(a + a')$ of G (which may also be written $[a(X - a'X_0) + a'(X' + aX_0)]/(a + a')$, the ratio on the left equals $L(X'')$ by linearity, while that on the right is not less than $Q(X'')$ by convexity and so, by hypothesis, not less than $L(X'')$. This completes the proof, which follows closely that given in a special case by Banach.

As a consequence of the theorem just proved, we deduce for our convex functionals $\varphi(W)$, still using only ideas set out by Banach,¹⁴ the following corollary:

(7.3) *Each of the functionals $\varphi(W)$ of §6 has a flat support at any W_0 for which its value $\varphi(W_0)$ is finite.*

To prove this, we write $Q(X)$ for the difference $\varphi(W_0 + X) - \varphi(W_0)$ and, since the latter vanishes for $X = 0$, we need only obtain a linear functional $L(X)$ nowhere exceeding $Q(X)$. We shall construct such an $L(X)$ by transfinite induction, first observing that it clearly exists in the space consisting of the single point 0. Let the points at which Q is defined and finite above be well-ordered, the origin 0 being first. Let X_0 be a subsequent point and let G be the vector subspace consisting of the linear combinations of vectors previous to X_0 .

We make the inductive hypothesis that $L(X)$ exists in G with the properties required. By (6.3) the functional Q is bounded above, and so finite above, on a segment containing 0 and X_0 in its interior, and therefore at some point of each of the sets $\{X_0, G\}_+$ and $\{X_0, G\}_-$. By (7.1) we can therefore continue $L(X)$ into $\{X_0, G\}$, and so, by transfinite induction, into the whole of the vector subspace consisting of points which are linear combinations of those at which the functional Q is defined and finite above. This subspace is precisely the whole of our vector space by (6.2).

8. Necessary conditions for an attained minimum

Denoting by U_0 a particular boundary-function, we now propose to examine some consequences of the hypothesis that one of the minima $\varphi(U_0)$ is attained, i.e. that for some surface S_0 having the boundary U_0 and belonging to the appropriate one of the four classes introduced in §5, we have $\varphi(U_0) = F(S_0)$.

We shall modify our original assumption concerning the function $f(x, y, p, q)$, and suppose that it is *not this function but the difference*

$$f(x, y, p, q) - f_0(x, y)$$

which is bounded above when x, y lies in A and p, q lies in a bounded set, where $f_0(x, y)$ denotes the average of f at x, y for the surface S_0 or, in particular, in the case of an ordinary surface, the value of f when we substitute for p, q the gradient of S_0 . In addition, we shall suppose, as is really implied in the fact that $F(S_0)$ exists, that $f_0(x, y)$ is integrable in some sense on every portion of A . It is then possible to write $f_0(x, y) = 0$ without effective loss of generality, and we shall do this when convenient in any proof.

We shall denote by $z_0(x, y)$ or simply z_0 , the track of S_0 , and by W_0 the abstract vector which specifies its boundary U_0 together with the discontinuity zero. We denote further by $z(x, y)$ or simply by z , the track of an arbitrary admissible Lipschitzian surface S , and by $\bar{f}(x, y)$ the average at x, y of the func-

¹⁴ Banach [1], p. 29, Corollary.

tion $f(x, y, p, q)$ for this surface. Finally we denote by Δ an arbitrary interval of the x, y plane.

(8.1) *There exists a linear functional $L(z; \Delta)$ of z , additive in Δ dependent only on the values of z on the boundary of $A \cdot \Delta$, such that*

$$(8.2) \quad \iint_{A \cdot \Delta} f(x, y) dx dy \geq \iint_{A \cdot \Delta} f_0(x, y) dx dy + L(z; \Delta) - L(z_0; \Delta)$$

for every admissible surface, (i.e. for every Lipschitzian surface of the appropriate class), whatever its boundary U may be.

(8.3) *In particular, taking for Δ an interval including A , there is a linear functional $L(z)$ depending only on the boundary-values U of the track z of the surface S , such that*

$$(8.4) \quad F(S) \geq F(S_0) + L(z) - L(z_0).$$

It will be observed that this last statement includes as a special case the minimizing hypothesis from which we started, namely that

$$F(S) \leq F(S_0) \quad \text{when} \quad U = U_0,$$

since we then have $L(z) = L(z_0)$. So that (8.3), and *a fortiori* (8.1), are theorems in which the hypotheses are exploited to the full.

We may suppose $f_0(x, y) = 0$, so that $f(x, y, p, q)$ now satisfies the hypotheses of §5. We have then $\varphi(W_0) = 0$, so that by (7.3) there is a linear functional of $\bar{W} - W_0$ nowhere exceeding $\varphi(\bar{W})$, and *a fortiori* not exceeding the values of $F(\bar{S})$ for a discontinuous \bar{S} of the appropriate type whose discontinuity and boundary are specified by \bar{W} . We select for \bar{S} a surface which coincides in Δ with an admissible continuous surface S and outside Δ with the minimizing surface S_0 . The linear functional of $\bar{W} - W_0$ obtained becomes a functional of the track of S and of the interval Δ , clearly dependent only on the boundary-values of this track z in $A \cdot \Delta$, and which we denote by $L(z - z_0, \Delta)$. The functional $L(z, \Delta)$ is moreover linear in z and additive in Δ , by construction; furthermore it satisfies (8.2) since the left-hand side is $F(\bar{S})$. This completes the proof.

We shall now proceed to derive from (8.2) further information regarding the functional $L(z, \Delta)$ whose existence is asserted there.

(8.5). *There exist bounded measurable functions of x, y , the functions $P(x, y)$ and $Q(x, y)$ say, such that*

$$(8.6) \quad L(z, \Delta) = \iint_{A \cdot \Delta} (P \cdot z_y - Q \cdot z_x) dx dy.$$

In proving this, we again suppose $f_0(x, y) = 0$. Moreover we may restrict ourselves, by homogeneity, to tracks z of surfaces whose Lipschitz constants are sufficiently small for $z + z_0$ to be the track of an admissible surface. The function $f(x, y, p, q)$ being now bounded above for the relevant surfaces, it

follows by applying (8.2) to surfaces with the track $z + z_0$, that $L(z, \Delta)$ cannot exceed $N \cdot |\Delta|$, where N is a constant and $|\Delta|$ is the area of Δ . We can therefore express $L(z, \Delta)$ as the integral of its derivative in Δ , the latter existing almost everywhere for fixed z .

In particular, when z is a linear function $ax + by + c$, the functional L , which is clearly unaltered by passing to a parallel surface, will be of the form $aA + bB$ where A and B are functions of Δ possessing bounded derivatives in Δ almost everywhere. Denoting by $-Q$ and $+P$ these derivatives at x, y , and writing for instance $P = Q = 0$ when they do not exist, the formula to be proved clearly holds for linear tracks z .

In order to establish the formula generally, we denote its right-hand side by $L_1(z, \Delta)$. We denote further by z_1 a track with the same boundary as z on $A \cdot \Delta$ and with a Lipschitz constant increased by as little as we please, whose gradient in a subset of Δ of measure as close as we please to Δ differs arbitrarily little from that of z and is constant in a neighborhood of each point of this subset.¹⁵ Now the difference of $L_1(z_1, \Delta)$ and $L_1(z, \Delta)$ is arbitrarily small since it is less than a constant multiple of the integral of the absolute difference of the gradients of z and z_1 . Moreover $L(z_1, \Delta)$ and $L_1(z_1, \Delta)$ have the same derivative outside small measure, and their derivatives elsewhere are bounded; therefore their difference also is as small as we please. Finally $L(z, \Delta)$ and $L(z_1, \Delta)$ coincide since z and z_1 have the same boundary on $A \cdot \Delta$. Combining these results, we see that $L(z, \Delta)$ and $L_1(z, \Delta)$ differ arbitrarily little, and so coincide. This completes the proof of (8.5).

These results can now be sharpened still further by applying the following lemma due to Haar¹⁶ and Schauder:¹⁷

(8.7) *In order that the integral $\iint_A (P \cdot z_y - Q \cdot z_x) dx dy$, where P, Q are bounded measurable functions of x, y , shall take the value 0 whenever z is a Lipschitzian function vanishing on the boundary of A , it is both necessary and sufficient that there exist a Lipschitzian function Z whose gradient is almost everywhere P, Q .*

It is assumed in the above lemma that A is a bounded simply-connected domain.¹⁸

As a consequence of these results, our main theorem becomes:

(8.8) *In order that $F(S_0) = \varphi(U_0)$, where S_0 satisfies the remaining hypotheses of the beginning of this paragraph, it is necessary and sufficient that there exist a Lipschitzian function Z such that the following inequality be verified for every admissible surface (whose track z has an arbitrary boundary):*

$$(8.9) \quad \iint_{A \cdot \Delta} \{f(x, y) - f_0(x, y) - Z_x(z_y - z_{0y}) + Z_y(z_x - z_{0x})\} dx dy \geq 0.$$

¹⁵ The existence of such a track is ensured by (12.1) p. 98 of the first note, Young [12].

¹⁶ Haar [2], [4].

¹⁷ Schauder [9].

¹⁸ Our results have been mainly stated for convex domains but hold generally. At this stage however they extend only to simply-connected domains.

9. Localized form of the conditions. Connection with Weierstrass' condition and with the Euler-Haar equations

Following Bonnet and Haar, we term *adjoint*¹⁹ of the minimizing surface S_0 , the ordinary surface with the track $Z(x, y)$ whose existence is established in (8.8).

Our object is to reduce (8.9) to the inequality

$$(9.1) \quad f(x, y, p, q) - f_0(x, y) - P \cdot (q - q_0) + Q \cdot (p - p_0) \geq 0,$$

where p_0, q_0 and P, Q are the gradients at x, y of the tracks z_0 and Z . The validity of (9.1) for all p, q will be termed *the condition* (W, E) at x, y ; similarly, its validity for all p, q interior to the circle of radius K and center at the origin, will be termed *the condition* $(W, E)_K$. These terms are similar to those adopted by the writer in the case of curves²⁰ and the most elementary considerations show that for a function $f(x, y, p, q)$ possessing partial derivatives in p and q at x, y (and in particular for the integrands of all classical problems), the condition (W, E) is wholly equivalent to the simultaneous validity of

$$f_{p_0} = -Q, \quad f_{q_0} = P,$$

$$f(x, y, p, q) - f_0(x, y) - f_{p_0} \cdot (p - p_0) - f_{q_0} \cdot (q - q_0) \geq 0.$$

The first two relations are Haar's form of the Euler equations, while the third is Weierstrass' condition.

A set E of values of x, y, p, q will be termed *superficially negligible*, if given any ordinary Lipschitzian surface S with the track $z(x, y)$, the set of the points x, y, p, q of E at which p, q is the gradient of this track has a projection in the x, y plane of measure zero. In view of this definition, it is clear that an alteration of $f(x, y, p, q)$ in such a set of values of x, y, p, q cannot affect the value of $F(S)$ for ordinary Lipschitzian surfaces. On the other hand, the example considered in §3 shows that such an alteration may affect the values of $F(S)$ for generalized S . This difference between the ordinary and the generalized problems will cause slight differences between the corresponding localized forms of the conditions for an attained minimum, which we state in (9.2) and (9.3) below.

We begin with the ordinary problem.

(9.2) *With the hypotheses of (8.8), where we now suppose S_0 an ordinary surface, a necessary and sufficient condition for an ordinary minimum is that the inequality (9.1) be verified for a corresponding adjoint surface with the gradient P, Q except at most in a superficially negligible set of x, y, p, q .*

This is clearly equivalent to the condition that the integrand of left-hand side of (8.9) be non-negative for every admissible surface except at most in a set of x, y of plane measure 0; and this in its turn is clearly equivalent to (8.9) itself. A similar result holds for the ordinary minimum in the class of surfaces with Lipschitz constants less than K .

¹⁹ Haar_A[5].

²⁰ Young_A[11].

The corresponding results for the generalized problem are more complete and also not so easy to deduce from (8.8).

(9.3) *With the hypotheses of (8.8), a necessary and sufficient condition for a minimum among generalized surfaces whose Lipschitz constants are either unrestricted or less than a fixed K , is that at almost points x, y of A the minimizing surface S_0 satisfy either the condition (W, E) , or else the condition $(W, E)_K$.*

The main difference between this statement and the one concerning ordinary minima is that more information is given here about the exceptional set of x, y, p, q for which (9.1) need not hold, its projection in the x, y plane being this time a null set. Incidentally the combination of the two results will show that if an ordinary minimum is attained, then by altering f for at most a superficially negligible set of x, y, p, q we ensure that the same minimizing surface provides a generalized minimum.

In proving (9.3), we may suppose as usual $f_0(x, y) = 0$. We shall denote by K_0 any number exceeding the Lipschitz constant of S_0 and less than the upper bound of the Lipschitz constants of all admissible surfaces. We shall denote further by $f^*(x, y, p, q)$ the lower bound of the expression

$$(9.4) \quad \sum_{k=1}^N a_k f(x, y, p_k, q_k)$$

as function of the integer N and of the a_k, p_k, q_k where $k = 1, \dots, N$ when these variables are restricted by the conditions

$$(9.5) \quad \begin{aligned} a_k &\geq 0, & p_k^2 + q_k^2 &\leq K_0^2, \\ \sum_{k=1}^N a_k &= 1, & \sum_{k=1}^N a_k p_k &= p, & \sum_{k=1}^N a_k q_k &= q. \end{aligned}$$

It is easy to see that the lower bound of (9.4) is unaltered if we impose the additional restriction $N \leq 3$. For if $N > 4$, there exists a system of numbers b_k , such that not all b_k vanish, which fulfil the equations

$$\sum_{k=1}^N b_k = \sum_{k=1}^N b_k p_k = \sum_{k=1}^N b_k q_k = 0;$$

we can arrange, by changing the signs if necessary throughout, that

$$\sum_{k=1}^N b_k f(x, y, p_k, q_k) \leq 0,$$

and, by multiplying through by a suitable positive constant (remembering that at least one b_k must be negative), that the *least* of the numbers

$$a'_k = a_k + b_k$$

is equal to zero. It follows that we can reduce the value of N without increasing (9.4), by taking as new system of a_k, p_k, q_k those of the a'_k, p_k, q_k for which a'_k does not vanish, and so, by induction we can make $N \leq 3$.

We shall require, for the purpose of proving (9.3), the following properties of the function $f^*(x, y, p, q)$:

- (9.6) (i) For constant x, y the function $f^*(x, y, p, q)$ is convex in p, q .
 (ii) For constant p, q the function $f^*(x, y, p, q)$ is measurable in x, y .

The first of these is an immediate consequence of the definition. To prove (ii), it is sufficient to prove that the set of x, y at which any given number c is not greater than f^* is measurable; this set is clearly the projection in the x, y plane of a set of x, y, p_k, q_k, a_k ($k = 1, 2, 3$) for which (9.5) holds and for which the expression (9.4) is greater than or equal to c ; as the projection of a set measurable B , it is therefore itself²¹ measurable (though not in general measurable B). This completes the proof of (9.6) (ii).

We shall still require one further lemma:

(9.7) Given any finite measurable function $h(x, y)$ greater than $f^*(x, y, p, q)$, there exists a generalized surface with constant gradient p, q and with a Lipschitz constant not exceeding K_0 , such that $M(f) < h$ for almost all x, y of A .

To prove this, we consider the Borel set of x, y, p_k, q_k, a_k ($k = 1, 2, 3$) in space of eleven dimensions, for which (9.5) holds and for which the expression (9.4) is less than $h(x, y)$. The projection of this set in the x, y plane contains every point of A , therefore²² the set itself contains the graph of a system of nine functions p_k, q_k, a_k ($k = 1, 2, 3$) of x, y measurable in A . Substituting these functions in (9.4), this expression becomes an average $M(f)$ whose value is a measurable function of x, y less than $h(x, y)$.

We are now in a position to establish (9.3). For any p, q of magnitude not exceeding K_0 , it is clear that f^* is a finite function of x, y not exceeding f , provided that we exclude if necessary a null set of x, y where $f = -\infty$ (if this last set were not a null set, it would still have positive measure when p, q is replaced by the gradient of S_0). Moreover, by excluding a suitable null set, we may suppose f^* , for the given constant p, q , to be measurable B in x, y . Choosing $h(x, y)$ to be $f^*(x, y, p, q) + \epsilon$ in this set, and say $f(x, y, p, q) + \epsilon$ in the complementary null set, it clearly follows from (9.7) and (8.9) that for almost all x, y of A ,

$$f^* + \epsilon \geq f_0 + Z_x(q - z_{0y}) - Z_y(p - z_{0x}).$$

From this inequality, making ϵ describe a sequence with limit 0 and making p, q describe all rational points of magnitude not exceeding K_0 , we deduce that outside a fixed null set, the inequality

$$f^* \geq f_0 + Z_x(q - z_{0y}) - Z_y(p - z_{0x})$$

holds for all such rational p, q . This same inequality, and *a fortiori* the inequality (9.1), then holds for the same set of x, y and all p, q whose magnitude

²¹ Lusin [7], p. 144 & p. 152; cf. also Saks [8], p. 47-50.

²² By the argument of Young [11], p. 237, lemma (5.2) & p. 256, 257.

does not exceed K_0 , since f^* is continuous in p, q on account of its finiteness and convexity. Finally, making K_0 describe a sequence with the limit ∞ or K , we see that (9.1) must still hold outside a null set of x, y for all relevant p, q , and this is what we asserted. The converse being a trivial consequence of (8.8), this completes the proof.

Let us remark in conclusion that the converse part of these theorems can be strengthened: A surface S_0 satisfying the condition (W, E) with a Lipschitzian adjoint, provides a minimum in a wider class of surfaces than those originally admitted, namely for all surfaces whose track z is such that the integral $\iint_A (Z_x z_y - Z_y z_x) dx dy$ depends only on the boundary values of the variable track z .

It has been proved in the case of a rectangle A , and the proof extends easily enough to the case in which A is a convex region, that the integral in question, where Z is a given Lipschitzian function, has this property for any z absolutely continuous in the sense of Tonelli.²³

CAPETOWN, SOUTH AFRICA.

REFERENCES

- [1] BANACH, S. *Théorie des opérations linéaires*, Monografie Matematyczne, I (Warszawa 1932).
- [2] HAAR, A. *Über die Variation der Doppelintegrale*, Journal für die reine und angewandte Mathematik, 146 (1919), pp. 1-18.
- [3] HAAR, A., *Über reguläre Variationsprobleme*, Acta Litterarum ac Scientiarum (Szeged), 3 (1927), pp. 224-234.
- [4] HAAR, A. *Über das Plateau'sche Problem*, Mathematische Annalen, 97 (1927), pp. 124-158.
- [5] HAAR, A. *Über adjungierte Variationsprobleme und adjungierte Extremalflächen*, Mathematische Annalen, 100 (1928), pp. 481-502.
- [6] HARDY, G. H., LITTLEWOOD, J. E., AND PÓLYA, G. *Inequalities*, (Cambridge 1934).
- [7] LUSIN, N. *Leçons sur les Ensembles Analytiques et leurs applications*, Collection de monographies Borel (Paris 1930).
- [8] SAKS, S. *Theory of the integral*, Monografie Matematyczne, VII (Warszawa 1937).
- [9] SCHAUDER, J. *Über die Umkehrung eines Satzes aus der Variationsrechnung*, Acta Litterarum ac Scientiarum (Szeged), 4 (1928), pp. 28-50.
- [10] STEINER, J. *Gesammelte Werke*, Band II.
- [11] YOUNG, L. C. *Necessary conditions in the calculus of variations*, Acta Mathematica, 69 (1938) p. 239-258.
- [12] YOUNG, L. C. *Generalized surfaces in the calculus of variations, I. Generalized Lipschitzian surfaces*. Annals of Mathematics, 43 (1942), pp. 84-103.
- [13] YOUNG, W. H. *On a formula for an area*, Proceedings of the London Mathematical Society, (2) 18 (1919), pp. 339-374.
- [14] YOUNG, W. H. *On a new set of conditions for a formula for an area*, Proceedings of the London Mathematical Society, (2) 21 (1923), pp. 75-94.

²³ Young, W. H. [13], [14]. The case in which the derivative in one variable of one of the functions Z, z and the derivative in the other variable of the other function have associated summabilities is also treated.

THE GEOMETRY OF ISOTROPIC SURFACES

BY SHIING-SHEN CHERN

(Received May 23, 1941; revised October 10, 1941)

Introduction

The geometry in a space in which there is given a family of varieties has been the object of extensive researches. The geometry of paths¹ corresponds to the case of a $(2n - 2)$ -parameter family of curves in an n -dimensional space, defined by a system of differential equations of the second order of a certain type. It may be regarded as a natural generalization of projective geometry, since Cartan has proved that of all the projective connections having the same geodesics there exists one, and only one, which is characterized by intrinsic properties and which he called normal.² Recently, M. Hachtroudi³ proved that in a three-dimensional space with a three-parameter family of surfaces there can be defined, in an intrinsic way, one and only one projective connection with the contact elements as the elements of the space. It is the aim of the present paper to study the geometry of a space in which there is given a two-parameter family of surfaces. Our main results may be summarized in the following two theorems:

THEOREM 1. *Suppose there be given in a three-dimensional space a two-parameter family of surfaces $\{\Sigma\}$ such that the tangent planes at a point of the surfaces of the family through the point do not pass through a fixed direction. Then there can be defined in the space one and only one four-dimensional Weyl geometry, which has the contact elements of the surfaces as its "points" and possesses certain intrinsic properties.*

THEOREM 2. *There exists a point transformation carrying one family of surfaces $\{\Sigma\}$ into another $\{\bar{\Sigma}\}$ when and only when the two four-dimensional Weyl geometries are equivalent.*

The paper is divided into four sections. In §1 we give the definition of a Weyl space of elements⁴ and of some of its fundamental notions. The determination of the Weyl space from the family of surfaces is then given in §2. In §3 we show that the so-defined Weyl geometry naturally leads to a solution of the problem of equivalence, i.e., the problem of deciding when two such families are equivalent under point transformations. The important particular

¹ Cf., for instance, O. Veblen and T. Y. Thomas, *The geometry of paths*, Trans. Amer. Math. Soc., vol. 25 (1923), pp. 551-608.

² E. Cartan, *Sur les variétés à connexion projective*, Bull. de la Soc. Math. de France, t. 52 (1924), pp. 205-241.

³ M. Hachtroudi, *Les espaces d'éléments à connexion projective normale*, Actualités Scientifiques et Industrielles, no. 565, Paris, 1937.

⁴ Cf. E. Cartan, *La méthode du repère mobile, la théorie des groupes continus, et les espaces généralisés*, Actualités Scientifiques et Industrielles, Paris, 1935.

case by which the Weyl space reduces to a three-dimensional point space is discussed in §4.

The extension of these results to a space of n -dimensions will be treated in a subsequent paper.

1. Weyl Space of Elements

Consider a three-dimensional space S with the coordinates x, y, z . To each point M of the space we attach three vectors I_1, I_2, I_3 , such that their scalar products satisfy the conditions

$$(1) \quad \begin{cases} I_1 I_3 = -I_2^2 \neq 0, \\ I_1^2 = I_3^2 = I_1 I_2 = I_2 I_3 = 0. \end{cases}$$

The totality of the point M and the vectors I_1, I_2, I_3 through M satisfying (1) will be called a *frame*. We say that the space is a Weyl space or that a Weyl connection is defined in the space, if a law of infinitesimal displacement of the following form is given:

$$\begin{cases} dM = \omega_1 I_1 + \omega_2 I_2 + \omega_3 I_3, \\ dI_1 = \omega_{11} I_1 + \omega_{12} I_2 + \omega_{13} I_3, \\ dI_2 = \omega_{21} I_1 + \omega_{22} I_2 + \omega_{23} I_3, \\ dI_3 = \omega_{31} I_1 + \omega_{32} I_2 + \omega_{33} I_3, \end{cases}$$

where the ω 's are Pfaffian forms in x, y, z . On differentiating (1) and making use of the above equations, we find

$$\begin{aligned} \omega_{13} &= 0, & \omega_{31} &= 0, & \omega_{11} + \omega_{33} - 2\omega_{22} &= 0, \\ \omega_{12} - \omega_{23} &= 0, & \omega_{21} - \omega_{32} &= 0. \end{aligned}$$

Hence it turns out to be more convenient to put

$$\tau = \omega_{12}, \quad \pi_1 = \omega_{11}, \quad \pi_2 = \omega_{22}, \quad \pi_3 = \omega_{21},$$

and to write the equations of infinitesimal displacement in the form:

$$(2) \quad \begin{cases} dM = \omega_1 I_1 + \omega_2 I_2 + \omega_3 I_3, \\ dI_1 = \pi_1 I_1 + \tau I_2, \\ dI_2 = \pi_3 I_1 + \pi_2 I_2 + \tau I_3, \\ dI_3 = \pi_3 I_2 + (-\pi_1 + 2\pi_2) I_3. \end{cases}$$

To serve our later purpose we shall generalize the notion of a Weyl space to that of a *Weyl space of elements*, by which we shall mean the following: We attach to each point of the space a one-parameter family of contact elements with

the point as origin. A contact element is then defined by four coordinates x, y, z, t , where x, y, z are the coordinates of the point and t fixes the plane. By attaching a frame to each contact element and supposing the $\omega_1, \omega_2, \omega_3, \tau, \pi_1, \pi_2, \pi_3$ in (2) to be Pfaffian forms in x, y, z, t , we shall get a Weyl space of elements. To define a Weyl space of elements it is thus necessary to give the one-parameter family of contact elements through each point and the seven Pfaffian forms in (2).

It will be advantageous to interpret the four-parameter family of contact elements to be the "points" of a four-dimensional auxiliary space S' . Given a curve in S' defined by the equations

$$(3) \quad x = x(\lambda), \quad y = y(\lambda), \quad z = z(\lambda), \quad t = t(\lambda),$$

we can express the Pfaffian forms in (2) as Pfaffian forms in λ . Then equations (2) become a system of ordinary differential equations. By a well-known theorem in the theory of differential equations, we see that when the initial frame $M^{(0)}I_1^{(0)}I_2^{(0)}I_3^{(0)}$ corresponding to an initial value $\lambda = \lambda_0$ is given, the family of frames $MI_1I_2I_3$ satisfying (2) is uniquely determined (with respect to $M^{(0)}I_1^{(0)}I_2^{(0)}I_3^{(0)}$). In this case, we say that *the family of frames is developed, along the curve (3), in the Euclidean space defined by the frame $M^{(0)}I_1^{(0)}I_2^{(0)}I_3^{(0)}$* . If we join a point P_0 of S' to a point P of S' by two different curves and develop, in the Euclidean space of the frame attached to P_0 , the two families of frames along these two curves, the frames obtained corresponding to the frame at P will in general be different.

Instead of studying the development of the frame at P on the frame at P_0 along two different curves, we shall consider an infinitesimal parallelogram with P_0 as a vertex and with two directions d and δ through P_0 as its two sides. Such a parallelogram is called a *cycle*. If we develop the family of frames first along the side d and then along δ or vice versa, the frames obtained will be different, so that there exists an infinitesimal transformation carrying one frame to the other. This infinitesimal transformation is called the *infinitesimal transformation associated to the cycle*. By a classical method,⁵ it is easy to show that the infinitesimal transformation associated to a cycle formed by the directions d and δ is given by equations of the form

$$(4) \quad \begin{cases} \nabla M = d\delta M - \delta dM = \Omega_1 I_1 + \Omega_2 I_2 + \Omega_3 I_3, \\ \nabla I_1 = d\delta I_1 - \delta dI_1 = \Pi_1 I_1 + T I_2, \\ \nabla I_2 = d\delta I_2 - \delta dI_2 = \Pi_3 I_1 + \Pi_2 I_2 + T I_3, \\ \nabla I_3 = d\delta I_3 - \delta dI_3 = \Pi_3 I_2 + (-\Pi_1 + 2\Pi_2) I_3, \end{cases}$$

⁵ Cf., for example, E. Cartan, *Leçons sur la géométrie des espaces de Riemann*, Paris, 1928, Chap. VII.

where, with the notations of Cartan's exterior calculus,⁶ we have

$$(5) \quad \begin{cases} \Omega_1 = \omega'_1 + [\pi_1 \omega_1] + [\pi_3 \omega_2], \\ \Omega_2 = \omega'_2 + [\tau \omega_1] + [\pi_2 \omega_2] + [\pi_3 \omega_3], \\ \Omega_3 = \omega'_3 + [\tau \omega_2] - [\pi_1 \omega_3] + 2[\pi_2 \omega_3], \\ T = \tau' - [\pi_1 \tau] + [\pi_2 \tau], \\ \Pi_1 = \pi'_1 + [\pi_3 \tau], \\ \Pi_2 = \pi'_2, \\ \Pi_3 = \pi'_3 + [\pi_1 \pi_3] - [\pi_2 \pi_3], \end{cases}$$

so that T, Ω_i, Π_i ($i = 1, 2, 3$) are exterior quadratic differential forms in x, y, z, t .

Suppose the frame attached to each contact element be so chosen that M coincides with the point of the element and I_1, I_2 belong to the plane of the element. Then the contact element will be fixed, if and only if

$$\omega_1 = \omega_2 = \omega_3 = \tau = 0.$$

It follows that $\omega_1, \omega_2, \omega_3, \tau$ are linear combinations of dx, dy, dz, dt and are themselves linearly independent. We can therefore put

$$(6) \quad \begin{cases} \Omega_i = P_{i1}[\omega_2 \omega_3] + P_{i2}[\omega_3 \omega_1] + P_{i3}[\omega_1 \omega_2] + \sum_{k=1}^3 Q_{ik}[\omega_k \tau], & i = 1, 2, 3, \\ \Pi_i = R_{i1}[\omega_2 \omega_3] + R_{i2}[\omega_3 \omega_1] + R_{i3}[\omega_1 \omega_2] + \sum_{k=1}^3 S_{ik}[\omega_k \tau], & i = 1, 2, 3, \\ T = P_1[\omega_2 \omega_3] + P_2[\omega_3 \omega_1] + P_3[\omega_1 \omega_2] + \sum_{k=1}^3 Q_k[\omega_k \tau]. \end{cases}$$

The 42 coefficients in (6) are the components of the "tensor of curvature and torsion" of the space.

The Weyl space of elements as defined above may be studied from two different points of view. We may either regard the space as a genuine four-dimensional space (x, y, z, t) and study its properties invariant under the group (or, more precisely, the pseudo-group) of transformations

$$(7) \quad \bar{x} = \bar{x}(x, y, z, t), \quad \bar{y} = \bar{y}(x, y, z, t), \quad \bar{z} = \bar{z}(x, y, z, t), \quad \bar{t} = \bar{t}(x, y, z, t),$$

or we may study the properties invariant under the group of transformations

$$(8) \quad \bar{x} = \bar{x}(x, y, z), \quad \bar{y} = \bar{y}(x, y, z), \quad \bar{z} = \bar{z}(x, y, z), \quad \bar{t} = \bar{t}(x, y, z, t).$$

⁶ An introduction to the notions of exterior multiplication and exterior differentiation is given in: E. Cartan, *Leçons sur les invariants intégraux*, Paris, 1922.

For our purpose we shall restrict ourselves to the latter point of view and shall call a property *intrinsic*, if it is invariant under transformations of the form (8). In particular, it follows that the concept of a point in S is an intrinsic property.

2. The Two-Parameter Family of Surfaces and Its Intrinsic Weyl Geometry

Suppose there be given in the space S a family $\{\Sigma\}$ of surfaces depending on two parameters. We shall study the properties of $\{\Sigma\}$, which are not affected by an arbitrary transformation of coordinates

$$(9) \quad \bar{x} = \bar{x}(x, y, z), \quad \bar{y} = \bar{y}(x, y, z), \quad \bar{z} = \bar{z}(x, y, z),$$

To make this study it will be convenient to introduce the four-dimensional auxiliary space S' formed by the contact elements of the surfaces of the family $\{\Sigma\}$. By the use of a parameter t the contact elements of $\{\Sigma\}$ having the same origin (x, y, z) can be defined by an equation of the form

$$(10) \quad \theta \equiv A(x, y, z, t)dx + B(x, y, z, t)dy + C(x, y, z, t)dz = 0,$$

where A, B, C are not all zero. The parameter t serves to distinguish the contact elements through the point and is arbitrary to the degree that it may be submitted to a transformation of the form

$$(11) \quad \bar{t} = \bar{t}(x, y, z, t).$$

In the auxiliary space S' of our contact elements with the coordinates x, y, z, t , each surface Σ of the family is represented by a two-dimensional variety. All these two-dimensional varieties fill up the space S' in the sense that through every point (in a certain neighborhood) of S' there passes one, and only one, such variety. The family $\{\Sigma\}$ of surfaces can therefore be defined by a completely integrable Pfaffian system of the form

$$(12) \quad \theta = 0, \quad \theta_1 \equiv A_1 dx + B_1 dy + C_1 dz + E dt = 0, \quad E \neq 0,$$

in the four variables x, y, z, t . The complete integrability of the system is expressed by the fact that the exterior derivatives θ', θ'_1 are congruent to zero, mod. θ, θ_1 .

By this representation of the family $\{\Sigma\}$ of surfaces in S as a family of surfaces in S' the properties of the original family unaffected by transformations of the form (9) are represented by properties of the corresponding family in S' , which remain unaltered under the transformations (9), (11), and conversely. But these are exactly the intrinsic properties in S' in the sense defined in §1.

Suppose, without loss of generality, that $C \neq 0$, so that we may simplify the system (12) to the form

$$(13) \quad \begin{cases} \theta \equiv dz - p dx - q dy = 0, \\ \theta_1 \equiv dt - r dx - s dy = 0, \end{cases}$$

where p, q, r, s are functions of x, y, z, t . We introduce the notation

$$(14) \quad \begin{cases} \frac{d}{dx} = \frac{\partial}{\partial x} + p \frac{\partial}{\partial z} + r \frac{\partial}{\partial t}, \\ \frac{d}{dy} = \frac{\partial}{\partial y} + q \frac{\partial}{\partial z} + s \frac{\partial}{\partial t}. \end{cases}$$

With this notation we have, for any function F in the variables x, y, z, t , the formula

$$(15) \quad dF = \frac{dF}{dx} dx + \frac{dF}{dy} dy + \frac{\partial F}{\partial z} \theta + \frac{\partial F}{\partial t} \theta_1,$$

and the integrability conditions of the system (13) can be written in the following simple form:

$$(16) \quad \frac{dp}{dy} - \frac{dq}{dx} = 0, \quad \frac{dr}{dy} - \frac{ds}{dx} = 0.$$

We now proceed to prove that when a two-parameter family of surfaces (13) is given in S such that

$$(17) \quad P \equiv p q_{tt} - q p_{tt} \neq 0,$$

there can be defined, in an intrinsic way, a Weyl geometry of elements in S . As it is easy to see, *the vanishing of P signifies that the contact elements through a point of S pass through a fixed direction.*

To each contact element formed by the point (x, y, z) and the plane $\theta = 0$ we attach a frame $MI_1I_2I_3$, adapted to it by supposing the following conditions to be satisfied:

a) M coincides with the point (x, y, z) .

b) I_1 is along the line of intersection of the plane and a neighboring plane through M , i.e., the line defined by the equations

$$(18) \quad \theta = 0, \quad p dx + q dy = 0.$$

We shall call this line the *characteristic* of the element.

c) I_2 lies on the plane $\theta = 0$.

If $MI_1I_2I_3$ is such a frame, the most general frame $M^*I_1^*I_2^*I_3^*$ satisfying the same conditions is given by

$$(19) \quad \begin{cases} M^* = M, & I_1 = u I_1^*, & I_2 = w I_2^* + \frac{uw}{w} I_1^*, \\ I_3 = \frac{w^2}{u} I_3^* + v I_2^* + \frac{1}{2} \frac{uw^2}{w^2} I_1^*, \end{cases}$$

where u, v, w are parameters, with $uw \neq 0$.

We now suppose our Weyl geometry to have the following two intrinsic properties:

1) Points are developed into points, i.e., the contact elements having the same origin are developed into the contact elements having the same origin.

2) The contact elements belonging to a fixed surface of the family are developed into contact elements having the same plane.

Let us express the conditions for these two properties analytically and see how far the Weyl geometry is thus defined. When the point M is fixed in space, we have

$$dx = dy = dz = 0.$$

Hence condition 1 is expressed by the relation

$$dM = 0 \pmod{dx, dy, dz},$$

i.e., $\omega_1, \omega_2, \omega_3$ are linear combinations of dx, dy, θ . Similarly, condition 2 is expressed by the fact that ω_3, τ are linear combinations of θ and θ_1 . It follows in particular that ω_3 is a multiple of θ . On the other hand, the condition b for the choice of the frame means that ω_2 is a linear combination of θ and $p_i dx + q_i dy$.

With a definite system of coordinates x, y, z, t we may make use of the parameters u, v, w in (19) to simplify the expressions for $\omega_1, \omega_2, \omega_3$. It is easy to verify that we may choose the frame attached to each contact element such that we have

$$(20) \quad \begin{cases} \omega_2 = -(p_i dx + q_i dy) + \eta \theta, \\ \omega_3 = \theta, \end{cases}$$

where η remains undetermined. The most general change of frame leaving the conditions a, b, c and the form of the equations (20) unaltered is given by

$$(21) \quad I_1 = I_1^*, \quad I_2 = I_2^* + v I_1^*, \quad I_3 = I_3^* + v I_2^* + \frac{1}{2} v^2 I_1^*,$$

v being a parameter.

In order to impose further conditions on the Weyl space, we are led to consider two particular kinds of cycles. The first kind is one whose directions d and δ displace the point of the contact element in the same direction and is analytically characterized by the conditions

$$\frac{\omega_1(d)}{\omega_1(\delta)} = \frac{\omega_2(d)}{\omega_2(\delta)} = \frac{\omega_3(d)}{\omega_3(\delta)},$$

or by

$$(22) \quad [\omega_1 \omega_2] = [\omega_2 \omega_3] = [\omega_3 \omega_1] = 0.$$

The second may be described as one displacing the surface of the contact element in the same direction and will be defined by

$$\frac{\omega_3(d)}{\omega_3(\delta)} = \frac{\tau(d)}{\tau(\delta)},$$

or by

$$(23) \quad [\omega_3 \tau] = 0.$$

The property of a cycle to be either one of the two kinds is intrinsic.

We now suppose our Weyl space to possess the following three further properties:

3) The infinitesimal transformation associated to every cycle displaces the point M in the direction of I_1 .

4) The infinitesimal transformation associated to every cycle whose directions d and δ displace the surface of the contact element in the same direction leaves the point M invariant.

5) The infinitesimal transformation associated to every cycle whose directions d and δ displace the point of the contact element in the same direction leaves the plane MI_1I_2 invariant.

We shall prove that the conditions 1 to 5 completely define the Weyl space.

To prove this, it is sufficient to determine the seven Pfaffian forms in (2). In the meantime we can still normalize the frame (with respect to the coordinates x, y, z, t) by making use of the transformation (21). Before doing this, we remark that the property 3 is expressed analytically by the fact that ∇M is equal to a multiple of I_1 , i.e., by the equations

$$(24) \quad \begin{cases} \omega'_2 + [\tau\omega_1] + [\pi_2\omega_2] + [\pi_3\omega_3] = 0, \\ \omega'_3 + [\tau\omega_2] - [\pi_1\omega_3] + 2[\pi_2\omega_3] = 0. \end{cases}$$

When these relations are satisfied, the property 4 is characterized by the condition that $\Omega_1 = 0$ when (23) holds, i.e., by

$$(25) \quad \omega'_1 + [\pi_1\omega_1] + [\pi_3\omega_2] + a[\tau\omega_3] = 0,$$

where a remains undetermined. Similarly, the property 5 is given by an equation of the form

$$(26) \quad \tau' - [\pi_1\tau] + [\pi_2\tau] - b[\omega_1\omega_3] - c[\omega_2\omega_3] = 0.$$

From the second equation of (24) we find, by making use of (20),

$$[(\tau + \theta_1)(p_idx + q_idy)] \equiv 0, \quad \text{mod } \theta.$$

Since τ is a linear combination of θ and θ_1 , it follows that τ is of the form

$$(27) \quad \tau = -\theta_1 + \xi\theta,$$

where ξ is undetermined. We calculate next the first equation of (24), mod ω_2, ω_3 . We find then

$$[(\omega_1 + p_idx + q_idy)\theta_1] \equiv 0, \quad \text{mod } \omega_2, \omega_3.$$

But ω_1 is a linear combination of dx, dy, θ only, so that it must be of the form

$$\omega_1 = -(p_idx + q_idy) + \sigma(p_idx + q_idy) + \xi\theta.$$

By applying the change of frame (21), we find that ω_1 is then changed into $\omega_1 + v\omega_2 + \frac{1}{2}v^2\theta$. We may therefore assume the frame so chosen that

$$(28) \quad \omega_1 = -(p_idx + q_idy) + \xi\theta.$$

When $\omega_1, \omega_2, \omega_3$ are of the forms given by (20), (28), the frame attached to each contact element is uniquely determined.

It now remains to determine the functions ξ, η, ζ and the Pfaffian forms π_1, π_2, π_3 from the conditions (24), (25), (26) and to show that they are thus completely determined. For this purpose we have to calculate the exterior derivatives of $\omega_1, \omega_2, \omega_3, \tau$. We remark that the calculation of every such

exterior derivative is equivalent to the calculation of the exterior derivative of a Pfaffian form of the form

$$(29) \quad \omega = l\theta - m dx - n dy,$$

where l, m, n are functions of x, y, z, t . We find, by making use of (20), (27), (28),

$$(30) \quad \left\{ \begin{aligned} P\omega' &= -\left(\frac{dm}{dy} - \frac{dn}{dx}\right)[\omega_1\omega_2] + \left\{\left(\frac{dm}{dy} - \frac{dn}{dx}\right)\eta + (m_t q_t - n_t p_t)\xi \right. \\ &\quad \left. + \left(\frac{dl}{dx} q_t - \frac{dl}{dy} p_t\right) + (m_x q_t - n_x p_t) + l(p_x q_t - q_x p_t)\right\}[\omega_1\omega_3] \\ &\quad + \left\{-\left(\frac{dm}{dy} - \frac{dn}{dx}\right)\xi + (-m_t q_{tt} + n_t p_{tt} - lP)\zeta + \left(-\frac{dl}{dx} q_{tt} + \frac{dl}{dy} p_{tt}\right) \right. \\ &\quad \left. + (-m_x q_{tt} + n_x p_{tt}) + l(-p_x q_{tt} + q_x p_{tt})\right\}[\omega_2\omega_3] \\ &\quad + (n_t p_t - m_t q_t)[\omega_1\tau] + (lP + m_t q_{tt} - n_t p_{tt})[\omega_2\tau] \\ &\quad + \{(m_t q_t - n_t p_t)\xi + (-m_t q_{tt} + n_t p_{tt} - lP)\eta + l_t P\}[\omega_3\tau]. \end{aligned} \right.$$

On substituting into (24), (25), (26), for $\omega'_1, \omega'_2, \omega'_3, \tau$ their expressions in terms of the exterior products of $\omega_1, \omega_2, \omega_3, \tau$, we shall get relations of the form

$$(31) \quad \left\{ \begin{aligned} -[\pi_1\omega_1] - [\pi_3\omega_2] &= \lambda_1[\omega_1\omega_2] + \lambda_2[\omega_1\omega_3] + \lambda_3[\omega_2\omega_3] + \lambda_4[\omega_1\tau] + \lambda_5[\omega_2\tau], \\ -[\pi_2\omega_2] - [\pi_3\omega_3] &= \mu_1[\omega_1\omega_2] + \mu_2[\omega_1\omega_3] + \mu_3[\omega_2\omega_3] + \mu_4[\omega_2\tau] + \mu_5[\omega_3\tau], \\ [\pi_1\omega_3] - 2[\pi_2\omega_3] &= \nu_1[\omega_1\omega_3] + \nu_2[\omega_2\omega_3] + \nu_3[\omega_3\tau], \\ [\pi_1\tau] - [\pi_2\tau] &= \rho_1[\omega_1\tau] + \rho_2[\omega_2\tau] + \rho_3[\omega_3\tau], \end{aligned} \right.$$

where we have, in particular,

$$(32) \quad \left\{ \begin{aligned} \lambda_1 &= \frac{1}{P} \left(\frac{dq_{tt}}{dx} - \frac{dP_{tt}}{dy} \right), & \lambda_4 &= \frac{1}{P} (p_t q_{tt} - q_t p_{tt}), \\ & & \lambda_5 &= \xi - \frac{1}{P} (p_{tt} q_{tt} - q_{tt} p_{tt}), \\ \mu_1 &= \frac{1}{P} \left(\frac{dq_t}{dx} - \frac{dp_t}{dy} \right), & \mu_4 &= \eta, & \mu_5 &= -\xi - \eta^2 + \eta_t, \\ \mu_2 &= \frac{1}{P} \left(\frac{d\eta}{dx} q_t - \frac{d\eta}{dy} p_t \right) + \frac{\eta}{P} \left(\frac{dp_t}{dy} - \frac{dq_t}{dx} + p_x q_t - q_x p_t \right) - \zeta \\ & & & + \frac{1}{P} (-p_t q_{tt} + q_t p_{tt}), \\ \nu_1 &= \frac{1}{P} (p_x q_t - q_x p_t), & \nu_2 &= \frac{1}{P} (-p_x q_{tt} + q_x p_{tt}) - \zeta, & \nu_3 &= -\eta, \\ \rho_1 &= \frac{1}{P} (r_t q_t - s_t p_t), & \rho_2 &= \zeta + \frac{1}{P} (-r_t q_{tt} + s_t p_{tt}). \end{aligned} \right.$$

In order that the system of equations (31) have a set of solutions for π_1, π_2, π_3 it is necessary and sufficient that the following relations hold:

$$(33) \quad \begin{cases} \lambda_1 - \mu_2 + \nu_2 - 2\rho_2 = 0, & \lambda_4 + \nu_3 - 2\mu_4 = 0, \\ \lambda_5 - \mu_5 = 0, & \mu_1 - \nu_1 + \rho_1 = 0. \end{cases}$$

The set of solutions is then given by

$$(34) \quad \begin{cases} \pi_1 = (-\nu_1 + 2\rho_1)\omega_1 + (-\nu_2 + 2\rho_2)\omega_2 + \lambda_2\omega_3 + (-\nu_3 + 2\mu_4)\tau, \\ \pi_2 = (-\nu_1 + \rho_1)\omega_1 + (-\nu_2 + \rho_2)\omega_2 + (\lambda_2 - \rho_3)\omega_3 + \mu_4\tau, \\ \pi_3 = -\mu_2\omega_1 + (\lambda_2 - \mu_3 - \rho_3)\omega_2 + \lambda_3\omega_3 + \mu_5\tau, \end{cases}$$

and is uniquely determined. Of the relations (33) the last one is identically satisfied, while the other three equations determine ξ, η, ζ . The Weyl connection is thus completely determined.

On summing up our results, we get the theorem:

Suppose that there is given in the space (x, y, z) a two-parameter family of surfaces such that the tangent planes to the surfaces through a point do not pass through a fixed direction. Then there can be defined in the space, in an intrinsic way, one and only one Weyl connection of elements having the following properties:

1. *Points are developed into points.*
2. *Surfaces of the family are developed into planes.*
3. *The infinitesimal transformation associated to every cycle displaces the point of the element in the direction of the characteristic of the element.*
4. *The infinitesimal transformation associated to every cycle by which the directions d and δ displace the surface of the element in the same "direction" leaves the element invariant.*
5. *The infinitesimal transformation associated to every cycle by which the directions d and δ displace the origin of the element in the same direction leaves the plane of the element invariant.*

3. The Problem of Equivalence

In the study of the geometry of a two-parameter family of surfaces in space under the group of transformations (9) the fundamental problem is the following: Given a two-parameter family $\{\Sigma\}$ of surfaces in S and another such family $\{\bar{\Sigma}\}$ in a space \bar{S} with the coordinates $\bar{x}, \bar{y}, \bar{z}$. When are they *equivalent*, i.e., when can the one be carried into the other by a transformation (9)? We shall show in this section that the solution of this problem follows as a consequence of the theorem of the last section, e.g., the possibility of defining an intrinsic Weyl connection in the space.

Before discussing this so-called "problem of equivalence," we shall give some further properties of the Weyl geometry defined in the last section. Our choice of the frames in §2 adapted to each contact element made use of the coordinates of the space. If we suppose only that the point M of the frame $MI_1I_2I_3$ coin-

cides with the point of the contact element and that the plane MI_1I_2 with the plane of the element, the frame attached to each contact element will depend on three parameters and the relation between two frames attached to the same element will be given by (19). By taking the parameters u, v, w in (19) as independent variables, we get a family of frames depending on the seven parameters x, y, z, t, u, v, w . Then the seven Pfaffian forms $\omega_1, \omega_2, \omega_3, \tau, \pi_1, \pi_2, \pi_3$ are linearly independent and it is easy to verify that they satisfy the system of equations (5), (6), which are called *the equations of structure of the Weyl space*. For this general family of frames the conditions

$$(35) \quad \Omega_1 = a[\omega_3\tau], \quad \Omega_2 = \Omega_3 = 0, \quad T = b[\omega_1\omega_3] + c[\omega_2\omega_3]$$

are still satisfied, because they have an intrinsic geometric meaning.

On account of these forms of the expressions for $\Omega_1, \Omega_2, \Omega_3, T$ the expressions for Π_1, Π_2, Π_3 can be simplified. In fact, by applying to the first four equations of (5) the theorem that the exterior derivative of the exterior derivative of a Pfaffian form is zero (the so-called "Poincaré's Theorem") and noticing that the resulting equations are identically satisfied if the expressions in (6) are zero, we get

$$(36) \quad \begin{cases} \Omega'_1 = [\Pi_1\omega_1] + [\Pi_3\omega_2] - [\Omega_1\pi_1] - [\Omega_2\pi_3], \\ \Omega'_2 = [T\omega_1] + [\Pi_2\omega_2] + [\Pi_3\omega_3] - [\Omega_1\tau] - [\Omega_2\pi_2] - [\Omega_3\pi_3], \\ \Omega'_3 = [T\omega_2] - [\Pi_1\omega_3] + 2[\Pi_2\omega_3] - [\Omega_2\tau] + [\Omega_3\pi_1] - 2[\Omega_3\pi_2], \\ T' = [T\pi_1] - [T\pi_2] - [\Pi_1\tau] + [\Pi_2\tau]. \end{cases}$$

These are some of the "Bianchi Identities" of the Weyl space. When the conditions (35) are satisfied, these equations can be simplified to the form

$$(37) \quad \begin{cases} [\Pi_1\omega_1] + [\Pi_3\omega_2] - [(da + 3a\pi_1 - 3a\pi_2)\omega_3\tau] = 0, \\ [\Pi_2\omega_2] + [\Pi_3\omega_3] + c[\omega_1\omega_2\omega_3] = 0, \\ [\Pi_1\omega_3] - 2[\Pi_2\omega_3] + b[\omega_1\omega_2\omega_3] = 0, \\ [\Pi_1\tau] - [\Pi_2\tau] - b[\omega_1\omega_2\tau] + [(b\omega_1 + c\omega_2)'\omega_3] + [(b\omega_1 + c\omega_2)\pi_2\omega_3] = 0. \end{cases}$$

From the third equation we see that every term of $\Pi_1 - 2\Pi_2 + b[\omega_1\omega_2]$ must contain ω_3 . On the other hand, we get, by multiplying the fourth equation by ω_3 ,

$$[(\Pi_1 - \Pi_2 - b\omega_1\omega_2)\tau\omega_3] = 0.$$

Similarly, the second and the third equations give respectively

$$[\Pi_2\omega_2\omega_3] = 0, \quad [(\Pi_1 - 2\Pi_2)\omega_2\omega_3] = 0,$$

from which it follows that

$$[(\Pi_1 - \Pi_2 - b\omega_1\omega_2)\omega_2\omega_3] = 0.$$

Again, by multiplying the right-hand sides of the first equation by ω_3 , the second by ω_2 , the third by $-\omega_1$, and adding, we get

$$[(\Pi_1 - \Pi_2 - b\omega_1\omega_2)\omega_1\omega_3] = 0.$$

It follows from the three equations that every term of $\Pi_1 - \Pi_2 - b[\omega_1\omega_2]$ must contain ω_3 . We can therefore put

$$(38) \quad \begin{cases} \Pi_1 = 3b[\omega_1\omega_2] + e[\omega_1\omega_3] + f[\omega_2\omega_3] + g[\omega_3\tau], \\ \Pi_2 = 2b[\omega_1\omega_2] + h[\omega_1\omega_3] + i[\omega_2\omega_3] + j[\omega_3\tau]. \end{cases}$$

From (37) it then follows that Π_3 is of the form

$$(39) \quad \Pi_3 = (h - c)[\omega_1\omega_2] + f[\omega_1\omega_3] + j[\omega_2\tau] + k[\omega_2\omega_3] + l[\omega_3\tau].$$

Consequently, for our Weyl space, the seven Pfaffian forms $\omega_1, \omega_2, \omega_3, \tau, \pi_1, \pi_2, \pi_3$ satisfy the equations of structure (5), with $\Omega_1, \Omega_2, \Omega_3, T, \Pi_1, \Pi_2, \Pi_3$ given by (35), (38), (39). In these equations there are introduced 11 quantities $a, b, c, e, f, g, h, i, j, k, l$, which are functions of x, y, z, t, u, v, w and which constitute the tensor of curvature and torsion of the space.

When two such Weyl spaces of elements are given, with the coordinates (x, y, z, t) and $(\bar{x}, \bar{y}, \bar{z}, \bar{t})$, we say that they are equivalent, if to every family of frames of the one (with one frame attached to each contact element), there is a corresponding family of frames of the other, such that a transformation of the form (7) exists, which satisfies the relations

$$(40) \quad \bar{\omega}_i = \omega_i, \quad \bar{\tau} = \tau, \quad \bar{\pi}_i = \pi_i, \quad i = 1, 2, 3.$$

If we attach to each contact element (x, y, z, t) the most general family of frames depending on three parameters u, v, w and to $(\bar{x}, \bar{y}, \bar{z}, \bar{t})$ the family of frames with the parameters $\bar{u}, \bar{v}, \bar{w}$, it is evident that the two Weyl spaces are equivalent, when and only when there exists a transformation of the form

$$(41) \quad \begin{cases} \bar{x} = \bar{x}(x, y, z, t, u, v, w), \dots, & \bar{t} = \bar{t}(x, y, z, t, u, v, w), \\ \bar{u} = \bar{u}(x, y, z, t, u, v, w), \dots, & \bar{w} = \bar{w}(x, y, z, t, u, v, w), \end{cases}$$

satisfying the relations (40).

With these preparations we now establish the theorem:

There exists a point transformation of the form (9) carrying one family of surfaces $\{\Sigma\}$ into another $\{\bar{\Sigma}\}$ when and only when the two corresponding Weyl spaces of elements are equivalent.

The proof of this theorem is immediate. In fact, when the two families of surfaces are equivalent, the corresponding Weyl spaces are equivalent, since our definition of the Weyl space is intrinsic. Conversely, when the Weyl spaces are equivalent, the transformation (41) will realize the equality of the sets of Pfaffian forms in (40). From the first three equations in (40), we see that in

the transformation (41) the coordinates $\bar{x}, \bar{y}, \bar{z}$ are functions of x, y, z alone, so that the transformation in question is a point transformation. From

$$\bar{\omega}_3 = \omega_3, \quad \bar{\tau} = \tau$$

we conclude that the transformation carries $\{\Sigma\}$ to $\{\bar{\Sigma}\}$. Hence the theorem is proved.

Analytically this result may be formulated as follows: *To solve the problem of equivalence for a two-parameter family of surfaces (whose tangent planes at a point do not pass through a fixed direction) in a three-dimensional space (x, y, z) , we introduce four auxiliary variables t, u, v, w and seven linearly independent Pfaffian forms ω_i, τ, π_i ($i = 1, 2, 3$) in the coordinates and in these variables. All these Pfaffian forms are invariant in the sense that the two families of surfaces $\{\Sigma\}$ and $\{\bar{\Sigma}\}$ are equivalent, when and only when there exists a transformation of the form (41) satisfying (40).*

Since the process of exterior differentiation is invariant under the point transformations (41), it follows that the 11 quantities a, \dots, l in (35), (38), (39) are invariants. We shall call them the *fundamental invariants* of the family $\{\Sigma\}$. From a given invariant F the equation

$$(42) \quad dF = F_1\omega_1 + F_2\omega_2 + F_3\omega_3 + F_0\tau + F'_1\pi_1 + F'_2\pi_2 + F'_3\pi_3$$

defines the new invariants $F_1, F_2, F_3, F_0, F'_1, F'_2, F'_3$, called the *covariant derivatives* of F . According to a theorem of E. Cartan,⁷ *the complete set of invariants of the family of surfaces consists of the fundamental invariants and their successive covariant derivatives*. This is to be understood as follows: By eliminating the independent variables x, y, z, t, u, v, w , we may set up relations of the form

$$(43) \quad \Phi(F_1, \dots, F_m) = 0$$

between the invariants F_1, \dots, F_m of our complete set. *Then the two families of surfaces $\{\Sigma\}$ and $\{\bar{\Sigma}\}$ are equivalent if and only if the relations (43) and*

$$\Phi(\bar{F}_1, \dots, \bar{F}_m) = 0$$

hold at the same time for F_1, \dots, F_m and for the corresponding invariants $\bar{F}_1, \dots, \bar{F}_m$ of $\{\bar{\Sigma}\}$. Moreover, it is also well-known that the equivalence of $\{\Sigma\}$ and $\{\bar{\Sigma}\}$ can be decided by only a finite number of such relations (43).

4. The Weyl Space of Points

An important particular case of our geometry is the case by which the infinitesimal transformation associated to every cycle depends only on the displacements of the origins of the contact elements. The space of elements is

⁷ E. Cartan, *Les sousgroupes des groupes continus de transformations*, Annales de l'Ecole Normale Supérieure, série 3, t. 25 (1908), pp. 57-194.

then identical with the three-dimensional point space (x, y, z) with a Weyl connection. The conditions for this are, from (35), (38), (39),

$$a = g = j = l = 0.$$

These conditions are, however, not independent. In fact, when $a = 0$, the first equation of (37) will give

$$g = l = 0.$$

It follows that *the conditions*

$$(44) \quad a = j = 0$$

are the necessary and sufficient conditions for the Weyl space to be a Weyl space of points.

We may give a geometrical interpretation to the condition $a = 0$. In fact, it signifies that the tangent planes to the surfaces of the family through a fixed point in space envelop a cone of the second order. To prove this, let δ be an operation under which the point remains fixed, so that

$$\omega_1(\delta) = \omega_2(\delta) = \omega_3(\delta) = 0.$$

Since t is now the only variable, we may determine the auxiliary variables u, v, w as functions of t such that the conditions

$$\pi_1(\delta) = \pi_2(\delta) = \pi_3(\delta) = 0$$

are also satisfied. We put $\tau(\delta) = e$, and we can assume, by changing the parameter t when necessary, that $\delta e = 0$.

If δ is an operation so chosen and d is an operation by which all the variables vary, we get, from (5) and (35),

$$(45) \quad \begin{cases} \delta\omega_1 = -ae\omega_3, \\ \delta\omega_2 = -e\omega_1, \\ \delta\omega_3 = -e\omega_2. \end{cases}$$

From (45) we get

$$(46) \quad \begin{cases} \delta^2\omega_3 = e^2\omega_1, \\ \delta^3\omega_3 = -ae^3\omega_3, \\ \delta^4\omega_3 = ae^4\omega_2 + (\dots)\omega_3, \\ \delta^5\omega_3 = -ae^5\omega_1 + (\dots)\omega_2 + (\dots)\omega_3. \end{cases}$$

If we define the "plane coordinates" l_1, l_2, l_3 of a plane through the point by the relation

$$(47) \quad l_1\omega_1 + l_2\omega_2 + l_3\omega_3 = 0,$$

we shall get, when the coordinates l_1, l_2, l_3 of the tangent plane of the family are expanded in powers of e ,

$$(48) \quad \begin{cases} l_1 = \frac{1}{2}e^2 - \frac{1}{120}ae^5 + \dots, \\ l_2 = -e + \frac{1}{24}ae^4 + \dots, \\ l_3 = 1 - \frac{1}{8}ae^3 + \dots. \end{cases}$$

From (48) it follows that

$$(49) \quad l_2^2 - 2l_1l_3 = \frac{1}{160}ae^5 + \dots.$$

Hence the cone

$$(50) \quad l_2^2 - 2l_1l_3 = 0$$

is the osculating quadric cone of the cone enveloped by the tangent planes of the family $\{\Sigma\}$ and the condition $a = 0$ is a necessary and sufficient condition that these tangent planes themselves envelop a quadric cone.

Added October 10, 1941. The author has succeeded in extending the results of this paper to a space of $n(\geq 3)$ dimensions. The following theorem has been established: *Given in a space of n dimensions a family of hypersurfaces depending on $n - 1$ parameters such that the tangent hyperplanes at a point to the hypersurfaces of the family through the point envelop a hypercone of $n - 1$ dimensions. Then it is possible to define in the space, in an intrinsic way, a Weyl geometry. The elements of this Weyl space are in general more complicated for $n \geq 4$ than for $n = 3$. The problem of equivalence is solved as a consequence of this result.*

NATIONAL TSING HUA UNIVERSITY
KUNMING, CHINA

IDEMPOTENT MARKOFF CHAINS

BY DAVID BLACKWELL

(Received March 6, 1942)

I. Introduction

Let \mathfrak{B} be any Borel field of subsets of an abstract space X , and suppose that for each $x \in X$ a probability measure¹ $P(x, E)$ is defined on \mathfrak{B} and that $P(x, E)$ is for fixed E a \mathfrak{B} -measurable function of x . Then $P(x, E)$ may be considered as representing the transition probability of going from the point x into the set E in a single trial, and it is said to determine a *Markoff chain* on X . The probability of going from x into E in n trials, denoted by $P_n(x, E)$, is given inductively by

$$(1) \quad P_n(x, E) = \int P(y, E) dP_{n-1}(x, y).^2$$

In this paper we shall consider Markoff chains for which $P_n(x, E)$ is independent of n . It is clear from (1) that this will occur whenever $P_2(x, E) \equiv P(x, E)$, so that we shall be studying Markoff chains which satisfy

$$(2) \quad P(x, E) = \int P(y, E) dP(x, y).$$

Such a Markoff chain will be called *idempotent*, and the justification for this is apparent.

Besides having some independent interest, idempotent Markoff chains are useful as tools in the study of general Markoff chains. For example if there is a subsequence $N_r \rightarrow \infty$ such that for all x and E

$$\frac{1}{N_r} \sum_{n=1}^{N_r} P_n(x, E) \rightarrow Q(x, E),$$

where $Q(x, E)$ is for each x a probability measure, then the Markoff chain determined by $Q(x, E)$ is idempotent, and the properties of $Q(x, E)$ are closely related to those of $P(x, E)$ (cf. Doob (II) and Yosida and Kakutani (III)).

Suppose that X contains only a finite number of points, denoted by the numbers $1, \dots, n$, and suppose that \mathfrak{B} consists of all subsets of X . Then $P(x, E)$ may be represented by a *Markoff matrix* $P = ||p_{ij}||$, i.e. a square matrix with non-negative elements and row sums equal to unity, where p_{ij} denotes the probability of going from i to j in a single trial. For this case (2) is the statement that the matrix P is idempotent. Idempotent Markoff matrices

¹ A probability measure is a non-negative completely additive set function having the value unity for the space.

² We shall use the notation $dP_n(x, y)$ to denote integration of y with respect to the measure $P_n(x, E)$.

have been completely characterized by Doob (II) and by Yosida and Kakutani (III), and the result is essentially summarized in the following theorem:

THEOREM 1: *If $P = || p_{ij} ||$ is an idempotent Markoff matrix, the subscripts may be divided into groups F, A_1, \dots, A_k such that*

$$(a) \quad p_{ij} = p_j > 0 \quad \text{for } i, j \in A_l$$

$$(b) \quad \sum_{j \in A_l} p_j = 1$$

$$(c) \quad p_{ij} \equiv 0 \quad \text{for } j \in F.$$

In section II we shall prove some general theorems about idempotent Markoff chains. In sections III and IV we shall apply these results to two classes of idempotent Markoff chains and show that in each case a decomposition analogous to that described by Theorem 1 obtains. Finally in section V we give a simple example of an idempotent Markoff chain which admits no such decomposition.

II. Definitions and general theorems

In this section we shall restrict attention to idempotent Markoff chains. A set $N \in \mathcal{B}$ is called a *null set* if $P(x, N) = 0$ for all x . A set $I \in \mathcal{B}$ is called *invariant* if there is a null set N such that

$$P(x, I) = 1 \quad \text{for all } x \in I - N.$$

The equation

$$(3) \quad \int_{\mathcal{B}} P(y, CE) dP(x, y) = 0 \quad \text{for all } x$$

is clearly a necessary and sufficient condition for the invariance of E . Following Doebelin (I), we shall call an invariant set *indecomposable* if it does not contain two disjoint non-null invariant subsets. A set E is called *strictly invariant* if $P(x, E) = 1$ for all $x \in E$. It is clear that any strictly invariant set is invariant and that if E is invariant there is a null set N such that $E - N$ is strictly invariant. We shall need the following fact about the reduction of multiple integrals:

LEMMA 1: *Let $\mathcal{B}(\mathcal{B}')$ be a Borel field of subsets of a space $Z(Y)$, and suppose that $P(y, E)$ is for each $y \in Y$ a measure on \mathcal{B} and for fixed E a \mathcal{B}' -measurable function of y . Let m be any measure on \mathcal{B}' , and define*

$$M(E) = \int P(y, E) dm.$$

Then any M -integrable function $f(z)$ is integrable with respect to $P(y, E)$ for all y except a set of m -measure zero, and

$$\int f(z) dM = \int \left\{ \int f(z) dP(y, z) \right\} dm.$$

PROOF: The verification of the lemma is immediate when $f(z)$ is the characteristic function of a set of finite M -measure, and the general result follows by approximation.

THEOREM 2: *If S, E are any sets, then for all x*

$$(4) \quad \int_S \left\{ \int_{CS} P(z, E) dP(y, z) \right\} dP(x, y) = \int_{CS} \left\{ \int_S P(z, E) dP(y, z) \right\} dP(x, y).$$

This theorem states that, starting from any point x , the probability of going first into S , then into CS , and finally into E is the same as the probability of going first into CS , then into S , and finally into E . This is not of course true for general Markoff chains, and in fact it may be shown that if it is true for every S in the special case $E = X$, then the Markoff chain is idempotent. The idea of the proof to be given below and of the results which follow will become clear at once if the integrals are interpreted as representing probabilities of events.

PROOF: For all x ,

$$(5) \quad \int_S \left\{ \int_{CS} P(z, E) dP(y, z) \right\} dP(x, y) + \int_{CS} \left\{ \int_{CS} P(z, E) dP(y, z) \right\} dP(x, y) \\ = \int \left\{ \int F(z) P(z, E) dP(y, z) \right\} dP(x, y),$$

where $F(z)$ is the characteristic function of CS . Also

$$(6) \quad \int_{CS} \left\{ \int_S P(z, E) dP(y, z) \right\} dP(x, y) + \int_{CS} \left\{ \int_{CS} P(z, E) dP(y, z) \right\} dP(x, y) \\ = \int_{CS} \left\{ \int P(z, E) dP(y, z) \right\} dP(x, y) = \int F(y) P(y, E) dP(x, y).$$

By Lemma 1, with $f(z) = F(z)P(z, E)$, the right members of (5) and (6) are equal. Hence their left members are equal, and (4) follows at once.

THEOREM 3: *The class \mathcal{I} of invariant sets is a Borel field.*

PROOF: It is clear that a denumerable sum of invariant sets is invariant; we need show only that the complement of an invariant set is also invariant. Taking $E = X$ in Theorem 2, (4) becomes

$$(7) \quad \int_S P(y, CS) dP(x, y) = \int_{CS} P(y, S) dP(x, y).$$

If S is invariant, it follows immediately from this equation and condition (3) that CS is also invariant.

THEOREM 4: *The function $P(x, E)$ is for fixed E an \mathcal{I} -measurable function.*

PROOF: We must show that the set

$$S = \{x \text{ for which } P(x, E) < k\}$$

is invariant for any k . We may assume $k > 0$. Now

$$(8) \quad \int_{CS} \left\{ \int_S P(z, E) dP(y, z) \right\} dP(x, y) \leq k \int_{CS} P(y, S) dP(x, y),$$

and the equality sign holds only if both sides vanish. Also

$$(9) \quad \int_S \left\{ \int_{CS} P(z, E) dP(y, z) \right\} dP(x, y) \geq k \int_S P(y, CS) dP(x, y).$$

By Theorem 2 the left members of (8) and (9) are equal, and (7) implies that their right members are equal. Hence equality holds in (8) for all x . Consequently the right side of (8) vanishes for all x and, since $k > 0$, this is precisely condition (3) for invariance of CS .

The principal result of this section is contained in the following theorem which shows that the strictly invariant indecomposable sets are the analogues of the sets A_i of Theorem 1.

THEOREM 5: *If S is strictly invariant and indecomposable, then $P(x, E)$ is independent of x for all $x \in S$.*

PROOF: For all $z \in S$ we have

$$(10) \quad P(z, E) = \int_S P(y, E) dP(z, y).$$

Let $y_0 \in S$. The sets

$$S \cdot \{P(x, E) \geq P(y_0, E)\} \quad \text{and} \quad S \cdot \{P(x, E) \leq P(y_0, E)\}$$

are by Theorem 4 invariant, and it is clear from (10) with $z = y_0$ that neither can be a null set. Since S is indecomposable, their complements in S , the sets

$$S \cdot \{P(x, E) < P(y_0, E)\} \quad \text{and} \quad S \cdot \{P(x, E) > P(y_0, E)\}$$

must be null sets. Then $P(x, E) = P(y_0, E)$ for all $x \in S$ except a null set, and (10) then implies that

$$P(x, E) = P(y_0, E) \quad \text{for all } x \in S.$$

III. The strictly separable case

A Borel field \mathcal{B} of subsets of X is said to be *strictly separable* if it is determined by a denumerable subcollection of its elements. \mathcal{B} is said to be *atomic* if $X = \sum_{\alpha} X_{\alpha}$, where each $X_{\alpha} \in \mathcal{B}$ and (11) for every $B \in \mathcal{B}$ and every α , either $BX_{\alpha} = X_{\alpha}$ or $BX_{\alpha} = O$. The sets X_{α} are called the *atoms* of \mathcal{B} .

LEMMA 2: *A strictly separable Borel field is atomic.*

This fact is well known; since we shall refer to the proof, we reproduce it here.

PROOF: Let \mathcal{B} be determined by F_1, F_2, \dots . Define $X_{\alpha} = \prod_{n=1}^{\infty} F_n^{\epsilon_n(\alpha)}$, where $\epsilon_n(\alpha) = \pm 1$, $E^1 = E$, $E^{-1} = CE$. Then each $X_{\alpha} \in \mathcal{B}$, $X = \sum_{\alpha} X_{\alpha}$, and the class of sets B satisfying (11) is a Borel field including every F_n and hence includes \mathcal{B} .

THEOREM 6: *If \mathcal{G} contains a strictly separable subfield \mathcal{G}_1 whose atoms are indecomposable, then $X = F + \sum_{\alpha} A_{\alpha}$, where*

$$(a) \quad P(x, E) = P_{\alpha}(E) \text{ for } x \in A_{\alpha}$$

$$(b) \quad P_{\alpha}(A_{\alpha}) = 1$$

$$(c) \quad F \text{ is a null set.}$$

PROOF: Let F_1, F_2, \dots determine \mathcal{G}_1 . We may assume the F_k form a field. Let N_k be the null subset of F_k for which $P(x, F_k) < 1$, and define $F = \sum_1^{\infty} N_k$. Let $\{X_{\alpha}\}$ be the atoms of I , and define $A_{\alpha} = X_{\alpha} - F$. If $x \in A_{\alpha}$, it is evident from the construction of the atoms given in the proof of Lemma 2 that $P(x, X_{\alpha}) = 1$ and hence that $P(x, A_{\alpha}) = 1$. Since A_{α} is strictly invariant and indecomposable, we may invoke Theorem 5 to obtain (a).

THEOREM 7: *If \mathcal{B} is strictly separable, the conclusion of Theorem 6 holds.*

PROOF: We shall define a subfield of \mathcal{G} satisfying the hypothesis of Theorem 6. Let F_1, F_2, \dots determine \mathcal{B} . We may suppose the F_j form a field. Define

$$I_{jkn} = \left\{ \frac{k}{n} \leq P(x, F_j) < \frac{k+1}{n} \right\}, \quad k = 0, \dots, n; n, j = 1, 2, \dots$$

Each of these sets is invariant by Theorem 4. Let \mathcal{G}_1 be the strictly separable subfield determined by the I_{jkn} . If X_{α} is an atom of \mathcal{G}_1 , it is clear that $P(x, F_j)$ is independent of x for all $x \in X_{\alpha}$ and all j . Since whenever a set of measures coincide on a field, they coincide on the Borel field determined by it, $P(x, E)$ is independent of x for all $x \in X_{\alpha}$ and all E . Then each X_{α} is indecomposable, for otherwise it would contain two disjoint non-empty strictly invariant subsets S and T , and $P(x, S) = 1$ for $x \in S$ and 0 for $x \in T$, contradicting the fact that $P(x, E)$ is independent of x on X_{α} .

Thus for instance if \mathcal{B} is the class of Borel sets in Euclidean space, any idempotent Markoff chain admits the decomposition of Theorem 6.

IV. The density case

Let m be any finite measure defined on \mathcal{B} and define the product measure $m \times m$ on the Borel field $\mathcal{B}_1 = \mathcal{B} \times \mathcal{B}$ of subsets of the product space $X \times Y$ of X with itself. Suppose $p_1(x, y)$ is any non-negative \mathcal{B}_1 -measurable function such that for each $x \in X$,

$$(12) \quad \int p_1(x, y) dy = 1.$$

Then the function

$$(13) \quad P(x, E) = \int_E p_1(x, y) dy$$

determines a Markoff chain on X , and the function $p_1(x, y)$ is called the *density function*. If the Markoff chain so defined is idempotent, we have

$$(14) \quad \int_{\mathbf{E}} p_1(x, y) dy = \int \left\{ \int_{\mathbf{E}} p_1(z, y) dy \right\} dP(x, z).$$

Applying a well known formula for transformation of measures and interchanging the order of integration, we obtain

$$(15) \quad \int_{\mathbf{E}} p_1(x, y) dy = \int_{\mathbf{E}} \left\{ \int p_1(x, z) p_1(z, y) dz \right\} dy;$$

i.e. if we define $p(x, y)$ by the equation

$$(16) \quad p(x, y) = \int p_1(x, z) p_1(z, y) dz,$$

$p(x, y)$ is another density function defining exactly the same Markoff chain as $p_1(x, y)$. It follows that for fixed x , $p(x, z)$ and $p_1(x, z)$ are equal for almost all z , and (16) then implies

$$(17) \quad p(x, y) = \int p(x, z) p_1(z, y) dz.$$

Multiplying both sides of (17) by $p_1(y, w)$ and integrating with respect to y , we obtain

$$(18) \quad p(x, w) = \int p(x, z) p(z, w) dz.$$

Thus if an idempotent Markoff chain can be represented by a density function, it can be represented by a density function which is idempotent in the sense of (18). (18) is the direct analogue of the idempotent Markoff matrix if we think of the Markoff chain as determined by the density function $p(x, y)$ rather than by the function $P(x, E)$. To study idempotent density functions we shall use the following well known fact:

LEMMA 3: If m is a finite measure defined on a Borel field \mathcal{B} of subsets of a space X , then

$$(19) \quad X = S + X_1 + X_2 + \dots \quad \text{where}$$

(a) $m(X_i) > 0$, and $E \in \mathcal{B}$, $E \subseteq X_i$ implies $m(E) = m(X_i)$ or $m(E) = 0$.

(b) every subset of S of positive measure has a subset of every smaller positive measure.

Either S or the X_i may of course be absent. The decomposition is clearly unique up to sets of measure zero.

PROOF: The sets X_i satisfying (a) may be divided into equivalence classes by defining two such sets as equivalent if their symmetric difference has measure zero. Any two sets in different classes can have only a set of measure zero in common, and since each set has positive measure there are at most denumerably

many equivalence classes. Let X_1, X_2, \dots be a set of representatives, one from each equivalence class, and define $S = X - \sum_1^\infty X_n$. Then every subset of S of positive measure has a subset of smaller positive measure and hence a subset of arbitrarily small positive measure, and we must show that this implies (b). Suppose $S_1 \subseteq S, 0 < k < m(S_1)$. Choose a sequence of sets K_n inductively so as to satisfy

$$(c) \quad K_n \in \mathcal{R}_n$$

$$(d) \quad m(K_n) > \text{l.u.b.}_{E \in \mathcal{R}_n} m(E) - \frac{1}{n}$$

where \mathcal{R}_n is the class of all subsets E of $S_1 - (K_1 + \dots + K_{n-1})$ for which $m(E) \leq k - m(K_1 + \dots + K_{n-1})$. Define $K = \sum_1^\infty K_n$. Then $K \subseteq S_1$ and $m(K) = k$, otherwise the choice of K_n would at some stage be contradicted, since $S_1 - K$ would contain a subset of positive measure smaller than $k - m(K)$.

We now use an argument similar to that of Doeblin in a related discussion (I). Apply Lemma 3 to m defined on the Borel field \mathcal{I} of invariant subsets of X , obtaining a decomposition (19) into invariant subsets. Each X_n is clearly indecomposable, and we may suppose that each X_n is strictly invariant. We shall show that S is a null set. First extract from S a null set H of largest possible measure and write $T = S - H$. It is sufficient to show that T has measure zero. Certainly every null subset of T has measure zero. Choose a class $D_{m,n} \in \mathcal{I}, n = 1, \dots, 2^m; m = 0, 1, 2, \dots$ of subsets of T so that

$$(a) \quad D_{mi} \cdot D_{mj} = O, \text{ if } i \neq j$$

$$(b) \quad m(D_{mi}) = \frac{1}{2^m} m(T)$$

$$(c) \quad D_{m+1,2i-1} + D_{m+1,2i} = D_{mi}.$$

By deleting a null set N_{mi} we shall make each of the D_{mi} strictly invariant. Hence by deleting $N = \sum_{m,i} N_{mi}$ from each D_{mi} simultaneously we obtain a new class of sets satisfying (a), (b), (c) and each of which is moreover strictly invariant. Thus without loss of generality we may suppose that each D_{mi} is strictly invariant. Now

$$D_{01} = \prod_{m=1}^{\infty} (D_{m1} + \dots + D_{m2^m}) = \sum_{i,j,k,\dots} D_{1i} \cdot D_{2j} \cdot D_{3k} \dots$$

Each set $D_{1i} \cdot D_{2j} \cdot D_{3k} \dots$ is a strictly invariant set of measure zero and is therefore empty. It follows that D_{01} is empty. Since $m(T) = m(D_{01})$, T has measure zero.

We are now in a position to apply Theorem 6, and we have shown further that there will be only denumerably many sets A_α in this case. However without appealing to this theorem we may easily obtain an even more precise decomposition in terms of the density function $p(x, y)$.

By Theorem 5, $P(x, E)$ is independent of x for $x \in X_n$; it follows that for any $x_1, x_2 \in X_n$,

$$(20) \quad p(x_1, y) = p(x_2, y)$$

except possibly on a y -set of measure zero. (18) then implies that (20) holds identically in y for all $x_1, x_2 \in X_n$. We write for all $x \in X_n$,

$$p(x, y) = p_n(y).$$

Then for any x it follows from (18) that

$$p(x, y) = \sum_{n=1}^{\infty} \int_{X_n} p(x, z) p_n(y) dz = \sum_{n=1}^{\infty} P(x, X_n) p_n(y);$$

i.e. any $p(x, y)$ is a linear combination of the functions $p_n(y)$. Define $U_n = \{y \in X_n \text{ such that } p_n(y) = 0\}$ and $V_n = \{y \in X_n \text{ such that } p_n(y) > 0\}$ and write $U = \sum_{n=1}^{\infty} U_n$, $V = \sum_{n=1}^{\infty} V_n$. Then $m(V) = 0$ and both U and V are null sets. Writing $A_n = X_n - (U + V)$, $F = X - (\sum_{n=1}^{\infty} X_n + V)$, we may summarize our results in the following theorem.

THEOREM 8: *If $p(x, y)$ is an idempotent density function, then $X = F + V + A_1 + A_2 + \dots$, where*

$$(a) \quad p(x, y) = p_n(y) \quad \text{for } x \in A_n$$

$$(b) \quad p_n(y) > 0 \quad \text{for } y \in A_n$$

$$(c) \quad \int_{A_n} p_n(y) dy = 1$$

$$(d) \quad p(x, y) \equiv 0 \quad \text{for } y \in F$$

$$(e) \quad m(V) = 0.$$

V. An example

We conclude by giving an example of an idempotent Markoff chain for which no decomposition like that obtained in the preceding sections is possible. Let \mathcal{B} be any Borel field, and define $P(x, E)$ to be the characteristic function of E . Then every set of \mathcal{B} is strictly invariant; there are no null sets except the empty set. A decomposition of X would consist in writing $X = \sum_{\alpha} X_{\alpha}$ where each X_{α} is indecomposable; and this is clearly impossible whenever \mathcal{B} is non-atomic.

INSTITUTE FOR ADVANCED STUDY

BIBLIOGRAPHY

- I. DOEBLIN, W., *Chaines simples constantes de Markoff*. Ann. Ec. Norm. (3), LVII., Fasc. 2.
- II. DOOB, J. L., *Topics in the theory of Markoff chains*. (To appear in the Trans. Am. Math. Soc.)
- III. YOSIDA, K., AND KAKUTANI, S., *Markoff processes with a denumerably infinite number of states*. Jap. Journal of Math., 1939, p. 47.

BANACH SPACE METHODS IN TOPOLOGY*

By SAMUEL EILENBERG

(Received January 22, 1942)

1. Introduction

We shall consider the totality $B(X)$ of all continuous real valued bounded functions f defined on a topological space X .¹ With the usual definition of addition and multiplication by real numbers and with the norm

$$\|f\| = \sup_{x \in X} |f(x)|$$

the set $B(X)$ becomes a Banach space.

Banach² has established the interesting and important result that two compact³ metric spaces X and Y are homeomorphic if and only if the corresponding spaces $B(X)$ and $B(Y)$ are isometric. Stone⁴ has shown that the hypothesis that X and Y are metric was superfluous. Both authors arrive at their results by establishing the general form of an isometric mapping of $B(X)$ onto $B(Y)$.

The Banach-Stone theorem implies that if X is compact then the topological structure of X is entirely determined by the metric properties of $B(X)$. The first purpose of this paper is to exhibit this relationship more explicitly. By studying the convex subsets of the surface of the unit sphere in $B(X)$ we define an "ideal space" Ξ and prove that if X is compact, then Ξ and X are homeomorphic. From this result we readily get a proof of the Banach-Stone theorem for compact spaces and also for a class of non-compact spaces including all metric spaces.

Although the space Ξ provides theoretically complete means for translating topological properties of a compact X into metric properties of $B(X)$, it will very often lead from simple topological properties of X to involved metric properties of $B(X)$ with no clear geometric interpretation. This makes it an interesting problem to relate "interesting" topological properties of X with "interesting" metric properties of $B(X)$. In this direction we prove that if $B(X)$ is a direct sum $B_1 + B_2$ of two Banach spaces, then $B_i = B(X_i)$ where X_i is an open and closed subset of X . From this we see that the connectedness properties of X translate into properties of $B(X)$ dealing with direct sums.⁵

* Presented to the American Mathematical Society April 12 and May 2, 1941.

¹ A *topological space* is a set X with a family of subsets called *open* such that (a) \emptyset and X are open, (b) the union of any number of open sets is open, (c) the intersection of two open sets is open. If in addition, (d) every two distinct points of X belong to two disjoint open sets, then X is said to be a *Hausdorff space*.

² S. Banach. *Théorie des opérations linéaires*. Warsaw, 1932, p. 170.

³ A space X is called *compact* if it is a Hausdorff space and if every covering of X by open sets contains a finite subcovering.

⁴ M. H. Stone. *Trans. Amer. Math. Soc.*, 41 (1937), p. 469.

⁵ Another result in this direction was recently announced by S. Krein and M. Krein (*C. R. U. R. S. S.*, 27 (1940), pp. 427-430): *If X is completely regular then $B(X)$ is separable and only if X is compact metric.*

$B(X)$ is not only a Banach space but also a linear lattice and a normed ring. From these two points of view $B(X)$ has been thoroughly investigated.⁶ In particular, necessary and sufficient conditions are known for a linear lattice or normed ring B in order that $B = B(X)$ for some topological space X . The analogous problem for a Banach space B is as yet unsolved and we hope that the methods developed in this paper will lead towards a solution.

2. Tychonoff's compacting

A Hausdorff space X is called *completely regular* if given a closed set $F \subset X$ and a point $x_0 \in X - F$ there is a continuous real valued function f on X such that $f(x_0) = 0$ and $f(x) = 1$ for all $x \in F$. Corresponding to every completely regular space X Tychonoff⁷ has constructed a space $\beta(X)$ such that

(T₁) $\beta(X)$ is compact,

(T₂) $X \subset \beta(X)$, $\bar{X} = \beta(X)$,

(T₃) every $f \in B(X)$ has an extension $f^* \in B[\beta(X)]$.

Moreover, these properties determine $\beta(X)$ uniquely. Since X is dense in $\beta(X)$ it is clear that the extension f^* of $f \in B(X)$ is unique and therefore establishes a 1 - 1 correspondence between $B(X)$ and $B[\beta(X)]$, in view of which the two Banach spaces may be considered identical.

The space $\beta(X)$ was extensively studied by Čech⁸ who has proved among other results that if X is completely regular and satisfies the first countability axiom⁹ then $\beta(X)$ determines X . More precisely, X is then the subset of $\beta(X)$ consisting of the points at which $\beta(X)$ has a countable base.⁹

3. Stars

Given a Banach space B we shall consider the transformation $Tb = -b$ and define

$$SA = A \cup TA$$

for every $A \subset B$.

LEMMA 3.1. Given $b_1, b_2 \in B$, $\|b_1\| \leq 1$, $\|b_2\| \leq 1$ the following conditions are equivalent:

(a) $\|tb_1 + (1 - t)b_2\| = 1$ for at least one t such that $0 < t < 1$,

(b) $\|tb_1 + (1 - t)b_2\| = 1$ for all t such that $0 \leq t \leq 1$,

(c) $\|b_1 + b_2\| = 2$.

⁶ See Stone *loc. cit.* and S. Kakutani, *Ann. of Math.*, 42 (1941), pp. 994-1024. The latter paper contains an up-to-date bibliography of the subject.

⁷ A. Tychonoff, *Math. Ann.*, 102 (1930), pp. 544-561.

⁸ E. Čech, *Ann. of Math.*, 38 (1937), pp. 823-844.

⁹ We say that a sequence U_i of open sets is a *countable base* for the point x of a Hausdorff space X if every open set containing x contains at least one of the sets U_i . If a countable base exists for every point of X , then X is said to satisfy the *first countability axiom*.

The proof is left to the reader.

For every $b \in B$ such that $\|b\| \leq 1$ we define the star $\text{St}(b)$ as follows:

$$\text{St}(b) = \{b_1 \mid \|b_1\| \leq 1, \|b_1 + b\| = 2\}.$$

The definition is justified geometrically by Lemma 3.1. We shall also consider the symmetrized star $\text{SSt}(b)$. Clearly the conditions $\|b\| < 1$, $\text{St}(b) = 0$ and $\text{SSt}(b) = 0$ are equivalent. Also the conditions $b_1 \in \text{St}(b_2)$ and $b_2 \in \text{St}(b_1)$ are equivalent and similarly for the symmetrized stars. It is also clear that $b \in \text{St}(b)$ if and only if $\|b\| = 1$.

Let now X be a topological space and $B(X)$ the Banach space described in the introduction.

LEMMA 3.2. *Let X be countably compact¹⁰ and let $f_1, f_2 \in B(X)$, $\|f_1\| \leq 1$, $\|f_2\| \leq 1$. We have $f_1 \in \text{St}(f_2)$ if and only if there is an $x \in X$ such that*

$$f_1(x) = f_2(x) = \pm 1.$$

Proof clear.

LEMMA 3.3. *Let X be a topological space and let $f \in B(X)$. We have $|f| = 1$ if and only if $\text{SSt}(f)$ is the whole surface of the unit sphere of $B(X)$.*

It is obvious that if $|f(x)| = 1$ for all $x \in X$ and $f_1 \in B(X)$ and $\|f_1\| = 1$ then either $\|f + f_1\| = 2$ or $\|f - f_1\| = 2$ and therefore $f_1 \in \text{SSt}(f)$.

Now suppose that $\inf |f(x)| = k < 1$. Define

$$f_1(x) = \frac{1 - |f(x)|}{1 - k}.$$

Clearly $f_1 \in B(X)$ and $\|f_1\| = 1$. However

$$\|f + f_1\| = \sup \left| \frac{f(x) - kf(x) + 1 - |f(x)|}{1 - k} \right| \leq 1 + k < 2$$

because if $f(x) \geq 0$ we have

$$|f(x) - kf(x) + 1 - |f(x)|| = |1 - kf(x)| \leq 1 - k^2$$

and if $f(x) \leq 0$ then

$$|f(x) - kf(x) + 1 - |f(x)|| = |1 - (2 - k)|f(x)|| \leq (1 - k)^2.$$

Similarly $\|f - f_1\| < 2$ and $f_1 \text{ non } \in \text{SSt}(f)$.

Lemma 3.2 shows that the elements of $B(X)$ such that $|f| = 1$ can be characterized intrinsically in $B(X)$. Hence the number of components of X can be determined by the properties of $B(X)$, since X will consist of n components if and only if there will be exactly 2^n continuous functions f such that $|f| = 1$.

4. The sets $M(f)$

Let $f \in B(X)$ and $\|f\| \leq 1$. Define

$$M(f) = \{x \mid |f(x)| = 1\}.$$

¹⁰ A space X is called countably compact if it is a Hausdorff space and if every countable covering of X by open sets contains a finite subcovering.

Clearly $M(f)$ is a closed subset of X . If $\|f\| < 1$ then obviously $M(f)$ is empty; if $\|f\| = 1$, $M(f)$ still may be empty unless X is countably compact.

LEMMA 4.1. *The sets $M(f)$ form a sublattice of the lattice of the closed subsets of X .*

In fact, we have $M(f_1) \cap M(f_2) = M(f)$ where $f(x) = \min \|f_1(x)|, |f_2(x)|\|$ and $M(f_1) \cup M(f_2) = M(f)$ where $f(x) = \max \|f_1(x)|, |f_2(x)|\|$.

LEMMA 4.2. *The space X is completely regular if and only if the sets $M(f)$ form a multiplicative base for the closed sets in X .*

If X is completely regular, then given a closed set $F \subset X$ and a point $x_0 \in X - F$, there is an $f \in B(X)$ such that $\|f\| \leq 1$, $f(x_0) = 0$ and $f(x) = 1$ for $x \in F$. Hence $F \subset M(f)$ and $x_0 \notin M(f)$. Conversely, if $M(f)$ form a multiplicative base, then for a given closed set $F \subset X$ and a point $x_0 \in X - F$ there is an $f \in B(X)$, such that $\|f\| \leq 1$, $F \subset M(f)$ and $x_0 \notin M(f)$. It follows that $|f(x_0)| < 1$ and $|f(x)| = 1$ for all $x \in F$; hence X is completely regular.

LEMMA 4.3.¹¹ *If X is normal then the sets $M(f)$ coincide with the closed G_δ subsets of X .*

LEMMA 4.4. *If X is countably compact then the relations*

$$M(f_1) \cap M(f_2) \neq 0 \quad \text{and} \quad f_2 \in \text{SSt}(f_1)$$

are equivalent for any $f_1, f_2 \in B(X)$ such that $\|f_1\| \leq 1$, $\|f_2\| \leq 1$.

In fact, $f_2 \in \text{SSt}(f_1)$ means that $\|f_1 + f_2\| = 2$ or that $\|f_1 - f_2\| = 2$. Since X is countably compact this means that for some $x_0 \in X$ either $|f_1(x_0) + f_2(x_0)| = 2$ or $|f_1(x_0) - f_2(x_0)| = 2$. This is equivalent with $|f_1(x_0)| = |f_2(x_0)| = 1$ which means $M(f_1) \cap M(f_2) \neq 0$.

LEMMA 4.5. *If X is completely regular and countably compact then the relations*

$$M(f_1) \subset M(f_2) \quad \text{and} \quad \text{SSt}(f_1) \subset \text{SSt}(f_2)$$

are equivalent for any $f_1, f_2 \in B(X)$ such that $\|f_1\| \leq 1$, $\|f_2\| \leq 1$.

Assume that $M(f_1) \subset M(f_2)$ and let $f \in \text{SSt}(f_1)$. In view of Lemma 4.4 we have then $M(f) \cap M(f_1) \neq 0$; consequently $M(f) \cap M(f_2) \neq 0$ and by Lemma 4.4 we have $f \in \text{SSt}(f_2)$. Hence $\text{SSt}(f_1) \subset \text{SSt}(f_2)$.

Let now $x_0 \in M(f_1) - M(f_2)$. Since X is completely regular there is an $f \in B(X)$ such that $\|f\| \leq 1$, $f(x_0) = 1$ and $f(x) = 0$ for all $x \in M(f_2)$. It follows that $x_0 \in M(f)$, therefore $M(f) \cap M(f_1) \neq 0$ and $f \in \text{SSt}(f_1)$. Since $f(x) = 0$ for all $x \in M(f_2)$ we have $|f(x) \pm f_2(x)| < 2$ and since X is countably compact this implies that $\|f \pm f_2\| < 2$. Consequently $f \in \text{SSt}(f_1) - \text{SSt}(f_2)$.

From Lemmas 4.3 and 4.5 we obtain the following

THEOREM 4.6. *If X is normal and countably compact, then the sets $\text{SSt}(f)$ ($f \in B(X)$, $\|f\| \leq 1$) ordered by inclusion form a lattice isomorphic with the lattice of the closed G_δ subsets of X .*

5. The sets $Q(x)$

Given a point $x \in X$ we define

$$Q(x) = \{f \mid f \in B(X), \|f\| = 1, |f(x)| = 1\}.$$

¹¹ For the proof see N. Vedenisoff, *Fund. Math.*, 27 (1936), pp. 234-238.

LEMMA 5.1. *Let X be completely regular. Given $x \in X$ and $f \in B(X)$ such that $\|f\| \leq 1$ we have $Q(x) \subset \text{SSt}(f)$ if and only if $|f(x)| = 1$.*

If $|f(x)| = 1$ it follows directly from the definition that $Q(x) \subset \text{SSt}(f)$. If now $|f(x)| < 1$ then there is an open set G such that $x \in G$ and $|f(x')| < 1 - \epsilon < 1$ for all $x' \in G$. Hence there is an $f_1 \in B(X)$ such that $\|f_1\| = 1$, $f_1(x) = 1$, and $f_1(x') = 0$ for all $x' \in X - G$. Consequently $f_1 \in Q(x)$ but since $\|f \pm f_1\| \leq 2 - \epsilon$ we have $f_1 \text{ non } \in \text{SSt}(f)$.

THEOREM 5.2. *If X is compact, then a subset A of the unit sphere of $B(X)$ is of the form $Q(x)$ for some $x \in X$ if and only if $A = SA_1$ where A_1 is a maximal convex subset of the surface of the unit sphere in $B(X)$.*

This is an immediate consequence of the following

THEOREM 5.3. *If X is compact then a subset A of the surface of the unit sphere of $B(X)$ is maximal convex if and only if there is an $x_0 \in X$ and $\epsilon = \pm 1$ such that $A = \{f \mid f \in B(X), \|f\| = 1, f(x_0) = \epsilon\}$.*

We first show that if such x_0 and ϵ exist then A is maximal convex. It is obvious that A is convex. Let then A^* be a convex set contained in the unit sphere of $B(X)$, such that $A \subset A^*$ and $f \in A^* - A$. Since $f(x_0) \neq \epsilon$ and X is compact and therefore normal there is an $f_1 \in B(X)$ such that $\|f_1\| = 1$, $f_1(x_0) = \epsilon$ and $f_1(x) = 0$ for $x \in M(f)$. It follows that $|f(x) + f_1(x)| < 2$ for all x and therefore $\frac{1}{2}(f_1 + f) \text{ non } \in A^*$. However, $f_1 \in A^*$ and $f \in A^*$, hence A^* was not convex.

In order to prove that every maximal convex subset A of the unit sphere is of the form described in Theorem 5.3, it is enough to find an $x_0 \in X$ and an $\epsilon = \pm 1$ such that $f(x_0) = \epsilon$ for all f in A . If we denote

$$M_+(f) = \{x \mid f(x) = -1\} \quad M_-(f) = \{x \mid f(x) = -1\}$$

this reduces to prove that either

$$\bigcap_{f \in A} M_+(f) \neq \emptyset \quad \text{or} \quad \bigcap_{f \in A} M_-(f) \neq \emptyset.$$

Assume the contrary. Since the sets $M_+(f)$ and $M_-(f)$ are closed and X is compact, we would have then two finite sequences $f_1, f_2, \dots, f_m, g_1, g_2, \dots, g_n$ in A such that

$$\bigcap M_+(f_i) = \emptyset \quad \text{and} \quad \bigcap M_-(g_i) = \emptyset.$$

Since A is convex we have $(f_1 + \dots + f_m + g_1 + \dots + g_n)/(n + m) \in A$, hence $\|f_1 + \dots + f_m + g_1 + \dots + g_n\| = n + m$, therefore there is an $x_1 \in X$ such that $f_1(x_1) = \dots = f_m(x_1) = g_1(x_1) = \dots = g_n(x_1) = \pm 1$ which leads to a contradiction.

6. Reconstruction of X from $B(X)$

Let B be a Banach space. A subset of B will be called an *ideal point*, and denoted by a Greek letter ξ if $\xi = SA$ where A is a maximal convex subset of the surface of the unit sphere of B . The totality of all the ideal points will be

noted by Ξ . A subset Φ of Ξ will be called an M -set if there is an $f \in B$ such that $\|f\| \leq 1$ and

$$\Phi = \{\xi \mid \xi \subset \text{SSt}(f)\}.$$

A subset of Ξ which is an intersection of M -sets will be called *closed*.

It is to be noted that with this definition of closed sets Ξ does not necessarily become a topological space. In fact, if Ξ were a topological space it would have to be a closed set and therefore an M -set. This would imply the existence of an element $f_1 \in B$ such that $\|f_1\| = 1$ and that $\Xi = \{\xi \mid \xi \subset \text{SSt}(f_1)\}$. From this it follows easily that $\text{SSt}(f_1)$ is the whole surface of the unit sphere of B . Of course, in most Banach spaces no such element exists.

THEOREM 6.1. *If X is a compact space then Ξ defined from the Banach space $B(X)$ is a topological space homeomorphic to X .*

Given $x \in X$ it follows from Theorem 5.2 that $Q(x)$ is an ideal point and that all the points of Ξ are obtained in this fashion. Hence taking

$$\xi(x) = Q(x)$$

we obtain a mapping of X onto Ξ .

We prove that ξ is 1-1. Let $x_1, x_2 \in X$ and $x_1 \neq x_2$. Since X is completely regular there is an $f \in B(X)$ such that $\|f\| \leq 1$, $f(x_1) = 1$ and $f(x_2) = 0$. Hence $f \in Q(x_1)$ but $f \notin Q(x_2)$, therefore $Q(x_1) \neq Q(x_2)$.

We now prove that ξ is a homeomorphism. Since the sets $M(f)$ form a multiplicative base for the closed sets in X (see Lemma 4.2) and by definition the M -sets form a similar base for the closed sets in Ξ , we shall prove that ξ is a homeomorphism by showing that if

$$\xi(F) = \Phi$$

then F is a set $M(f)$ if and only if Φ is an M -set.

Suppose that $F = M(f)$. It follows from Lemma 5.1 that $Q(x) \subset \text{SSt}(f)$ if and only if $|f(x)| = 1$, i.e. if $x \in F$. Consequently

$$\Phi = \{\xi \mid \xi \subset \text{SSt}(f)\}$$

and Φ is an M -set.

If f will vary over all of the unit sphere of $B(X)$, F will vary over all $M(f)$ sets and $\xi(F) = \Phi$ will vary over all M -set in Ξ . Since ξ is 1-1, the proof is complete.

From Theorem 6.1 and from §2, we obtain

THEOREM 6.2. *If X is a completely regular space, then Ξ defined from the Banach space $B(X)$ is a topological space homeomorphic with the compacting $\beta(X)$ of X .*

If, in addition, X satisfies the first countability axiom, then X is homeomorphic with the subset of Ξ consisting of points which have a countable base.

7. Isometric mappings of $B(X)$ onto $B(Y)$

THEOREM 7.1. *Given two topological spaces X and Y , every isometric mapping T of $B(X)$ onto $B(Y)$ is of the form*

$$(*) \quad Tf = g_0 + g_1 \cdot (T'f)$$

where $g_0, g_1 \in B(Y)$, $|g_1| = 1$ and T' is an isometry of $B(X)$ onto $B(Y)$ that is an algebraic and lattice isomorphism.

Define $T_1f = Tf - T0$. Clearly T_1 is an isometry and $T_10 = 0$; hence by a theorem of S. Mazur and S. Ulam¹² T_1 and its inverse are linear. Consequently we have $T_1\text{St}(f) = \text{St}(T_1f)$. Consider the function 1 in $B(X)$, since $\text{SSt}(1)$ is the whole surface of the unit sphere it follows that $\text{SSt}(T_11)$ is the whole surface of the unit sphere in $B(Y)$ and by Lemma 3.3 $|T_11| = 1$. If we define $g_0 = T0$, $g_1 = T_11$ and $T'f = T_1f/T_11$ we have $(*)$ and T' is a linear isometry such that $T'0 = 0$ and $T'1 = 1$. All that remains to be proved is that T' preserves the lattice properties.

Suppose $f \geq 0$, then¹³

$$\begin{aligned} T'f &= T' \|f\| + T'f - T' \|f\| = \|f\| - T'(\|f\| - f) \\ &\geq \|f\| - \|T'(\|f\| - f)\| = \|f\| - \|\|f\| - f\| \geq 0 \end{aligned}$$

and $T'f \geq 0$. The same holds for the inverse of T' and therefore T' is a lattice isomorphism.

THEOREM 7.2. *Let X and Y be two compact spaces. Every isometric mapping T of $B(X)$ onto $B(Y)$ is of the form*

$$Tf = g_0 + g_1 \cdot (fh)$$

where $g_0, g_1 \in B(Y)$, $|g_1| = 1$ and h is a homeomorphism $h(Y) = X$.

If $T0 = 0$ then $g_0 = 0$ and

$$Tf = g_1 \cdot (fh).$$

If $T0 = 0$ and $T1 = 1$, then $g_0 = 0$, $g_1 = 1$ and

$$Tf = fh.$$

The same holds if X and Y are completely regular and satisfy the first countability axiom; in particular, if X and Y are metric spaces.

In view of the previous theorem we may restrict ourselves to the case when $T0 = 0$ and $T1 = 1$. The isometry T is then linear and preserves all the lattice properties. In particular

$$(**) \quad T|f| = |Tf|.$$

We first discuss the case when X and Y are compact. Let Ξ_X and Ξ_Y be the corresponding ideal spaces and ξ_X and ξ_Y the homeomorphisms $\xi_X(X) = \Xi_X$

¹² C. R. Paris 194 (1932), pp. 946-948; also Banach *loc. cit.* p. 168.

¹³ A real number a is identified here with the function of $f \in B(X)$ such that $f(x) = a$ for all $x \in X$.

and $\xi_r(Y) = \Xi_r$. Since Ξ_x and Ξ_r were defined intrinsically using $B(X)$ and $B(Y)$ it is clear that T induces a homeomorphism between Ξ_x and Ξ_r which we will also denote by T ; $T(\Xi_x) = \Xi_r$. Define

$$h(y) = \xi_x^{-1} T^{-1} \xi_r(y) \quad \text{for } y \in Y.$$

Clearly h is a homeomorphism $h(Y) = X$. It remains to prove that $Tf = fh$ for all $f \in B(X)$.

We shall first prove that

$$(**) \quad \text{if } \|f\| \leq 1 \text{ then } M(f) = h[M(Tf)].$$

This follows immediately from the remark that $\xi_x[M(f)]$ and $\xi_r[M(Tf)]$ are the M -sets

$$\xi_x[M(f)] = \{\xi \mid \xi \subset \text{SSt}(f)\}$$

$$\xi_r[M(Tf)] = \{\xi \mid \xi \subset \text{SSt}(Tf)\} = \{\xi \mid \xi \subset T\text{SSt}(f)\},$$

hence $T\xi_x[M(f)] = \xi_r[M(Tf)]$ which implies $(**)$ in view of the definition of h .

Now let $f \in B(X)$, $x \in X$, $h(y) = x$ and $f(x) = a$. Consider

$$f_1 = 1 - \frac{|f - a|}{\|f - a\|}.$$

Because of $(**)$ we have

$$Tf_1 = 1 - \frac{|Tf - a|}{\|Tf - a\|}.$$

Since $x \in M(f_1)$ we shall have in view of $(**)$ $y \in M(Tf_1)$, hence $|(Tf_1)(y)| = 1$ which implies $(Tf)(y) = a$. This shows that $Tf = hf$ which proves the theorem in the compact case.

If now X and Y are completely regular, we consider the compactings $\beta(X)$ and $\beta(Y)$. Since $B[\beta(X)] = B(X)$ and $B[\beta(Y)] = B(Y)$ we have an isometry T mapping $B[\beta(X)]$ onto $B[\beta(Y)]$ and hence a homeomorphism $h[\beta(Y)] = \beta(X)$ such that $Tf = fh$. If X and Y satisfy the first countability axiom, then they are the sets of points at which the first countability axiom holds in $\beta(X)$ and $\beta(Y)$ respectively. Therefore $h(Y) = X$.

8. Discussion of Banach's proof

Theorem 7.2 was first established by Banach² for X and Y compact metric. His proof is based on the concept of a peak function. A function $f \in B(X)$ is called a *peak function* with the point $x_0 \in X$ as *peak* if $|f(x)| < |f(x_0)|$ for all $x \in X$, $x \neq x_0$. Banach proves it if X is compact metric, then $f \in B(X)$ is a peak function if and only if $f \neq 0$ and the limit

$$\lim_{t \rightarrow 0} \frac{\|f + tf_1\| - \|f\|}{t}$$

exists for every $f_1 \in B(X)$. Having such an intrinsic characterization of peak functions, it is easy to see that an isometry $TB(X) = B(Y)$ such that $T0 = 0$ must carry peak function into peak functions and this way a homeomorphism between X and Y can readily be established.

We propose to replace Banach's characterization of peak functions by the following one which seems to be more geometrical and natural:

THEOREM 8.1. *Let X be completely regular and countably compact. A function $f \in B(X)$ such that $\|f\| = 1$ is a peak function if and only if $\text{St}(f)$ is convex.*

If f is a peak function then there is an $x_0 \in X$ such that

$$|f(x_0)| = 1, \quad |f(x)| < 1 \quad \text{for } x \neq x_0.$$

Since X is countably compact it follows from Lemma 3.2 that

$$\text{St}(f) = \{f_1 \mid f_1 \in B(X), \|f_1\| = 1, f_1(x_0) = f(x_0)\}$$

and this set is clearly convex.

If f is not a peak function, then since X is countably compact f assumes its maximum and therefore there are at least two distinct points $x_1, x_2 \in X$ such that

$$|f(x_1)| = |f(x_2)| = 1.$$

Since X is completely regular there are two open sets G_1 and G_2 such that

$$x_1 \in G_1, \quad x_2 \in G_2 \quad \text{and} \quad G_1 \cap G_2 = \emptyset$$

and two functions $f_1, f_2 \in B(X)$ such that $\|f_1\| = \|f_2\| = 1$ and

$$f_i(x_i) = f(x_i), \quad f_i(x) = 0 \quad \text{for } x \in X - G_i, i = 1, 2.$$

Clearly $\|f + f_1\| = 2, \|f + f_2\| = 2$ and $\|f_1 + f_2\| = 1$ since $X = (X - G_1) \cup (X - G_2)$. Hence $f_1 \in \text{St}(f), f_2 \in \text{St}(f)$ and $\frac{1}{2}(f_1 + f_2) \notin \text{St}(f)$. Hence $\text{St}(f)$ is not convex.

Having an intrinsic characterization of the peak functions we can carry out Banach's proof provided we know that for every $x_0 \in X$ there is a peak function with x_0 as peak. It can be easily seen that this is true if and only if X is completely regular and every point of X is a G_δ . If in addition X is countably compact, this last condition becomes equivalent with the first countability axiom. Hence we see that Banach's proof can be carried out for completely regular, countably compact spaces satisfying the first countability axiom.

9. Representations of $B(X)$ as a direct sum

Given two linear subspaces B_1, B_2 of a Banach space B we say that B is the direct sum

$$B = B_1 + B_2$$

if every $b \in B$ admits a unique decomposition

$$b = b_1 + b_2, \quad b_1 \in B_1, \quad b_2 \in B_2$$

and if¹⁴

$$\|b\| = \max(\|b_1\|, \|b_2\|).$$

It is easy to prove that if $B = B_1 + B_2$ then B_1 and B_2 are closed subspaces of B .

Let X be a topological space and let

$$(*) \quad X_1 = X_1 \cup X_2, \quad X_1 \cap X_2 = 0, \quad \bar{X}_1 = X_1, \quad \bar{X}_2 = X_2.$$

Let $B(X_1), B(X_2)$ be the subspaces of $B(X)$ defined as follows

$$B(X_i) = \{f \mid f(x) = 0 \text{ for all } x \text{ non } \in X_i\}, \quad i = 1, 2.$$

We verify easily that $B(X) = B(X_1) + B(X_2)$.

LEMMA 9.1. *Every decomposition $X = X_1 \cup X_2$ of X into two disjoint closed sets determines a direct sum decomposition $B(X) = B(X_1) + B(X_2)$.*

We shall show that the converse is also true.

THEOREM 9.2. *Let X be a topological space. For every decomposition*

$$B(X) = B_1 + B_2$$

as a direct sum there is a decomposition

$$X = X_1 \cup X_2$$

into disjoint closed sets such that $B_1 = B(X_1)$ and $B_2 = B(X_2)$.

We first assume that X is compact. Since $B(X) = B_1 + B_2$ every $f \in B(X)$ has a unique decomposition

$$f = f_1 + f_2, \quad f_1 \in B_1, \quad f_2 \in B_2.$$

Define

$$Tf = f_1 - f_2.$$

Clearly T is linear and

$$\|Tf\| = \|f\|,$$

hence $TB(X) = B(X)$ is an isometry such that $T0 = 0$ and because of Theorem 7.2 we have

$$Tf = g_1 \cdot (fh)$$

where $g_1 \in B(X)$, $|g_1| = 1$, and h is a homeomorphic mapping of X onto itself.

We shall prove that $h(x) = x$ for all $x \in X$. Assume to the contrary that $h(x_0) = x_1 \neq x_0$ for some $x_0 \in X$. There is then an open set $G \subset X$ such that

$$x_0 \in G, \quad G \cap h(G) = 0.$$

¹⁴ The definition of the direct sum essentially depends upon the expression chosen for the norm. The most commonly used are $\|b\| = (\|b_1\|^p + \|b_2\|^p)^{1/p}$ for $p \geq 1$. The expression in the text is obtained by letting $p \rightarrow \infty$.

Since X is compact it is also completely regular and hence there is an $f \in B(X)$ such that

$$f(x_0) = 1, \quad f(x) = 0 \text{ for all } x \in X - G, \quad \|f\| = 1.$$

Since $f = f_1 + f_2$ where $f_1 \in B_1$, $f_2 \in B_2$ and $\|f\| = \max(\|f_1\|, \|f_2\|)$ we may assume that $\|f_1\| = 1$. Since X is compact there is an $x' \in X$ such that

$$|f_1(x')| = 1.$$

Since

$$|f_1(x') + f_2(x')| = |f(x')| \leq \|f\| = 1$$

$$|f_1(x') - f_2(x')| = |Tf(x')| \leq \|Tf\| = 1$$

we must have $f_2(x') = 0$. Consequently $f(x') = f_1(x') = \pm 1$ and therefore $x' \in G$. Since $f_1 \in B_1$ we have $Tf_1 = f_1$ and therefore

$$f_1 = g_1 \cdot (f_1 h).$$

Consequently $|f_1(h(x'))| = |f_1(x')| = 1$ and by the previous argument $h(x') \in G$. Hence $G \cap h(G) \neq \emptyset$, a contradiction.

Having proved that $h(x) = x$ for all $x \in X$ we have

$$Tf = g_1 \cdot f \quad \text{where} \quad |g_1| = 1.$$

Define

$$X_1 = \{x \mid g_1(x) = 1\}, \quad X_2 = \{x \mid g_1(x) = -1\}.$$

Clearly X_1, X_2 are closed and $X = X_1 \cup X_2$. Given $f \in B(X)$ we have $f \in B_1$ if and only if $Tf = f$, this means that $f = g_1 \cdot f$ and this holds if and only if $f(x) = 0$ whenever $g_1(x) = -1$. Hence we see that $B_1 = B(X_1)$. Analogously $B_2 = B(X_2)$.

Next we assume that X is completely regular. Let $\beta(X)$ be the compacting of X . Since $B(\beta(X)) = B(X)$ we have a direct sum decomposition $B(\beta(X)) = B_1 + B_2$ and hence there is a decomposition

$$\beta(X) = Y_1 \cup Y_2$$

into disjoint closed sets such that $B_1 = B(Y_1)$ and $B_2 = B(Y_2)$. Let

$$X_1 = X \cap Y_1, \quad X_2 = X \cap Y_2.$$

From the definition of the compacting it is clear that $Y_1 = \beta(X_1)$ and $Y_2 = \beta(X_2)$. Hence $B_1 = B(X_1)$, $B_2 = B(X_2)$.

If now X is an arbitrary topological space we consider the completely regular space $\rho(X)$ defined by Čech.⁸ Again $B(\rho(X)) = B(X)$ and hence we have the direct sum decomposition $B(\rho(X)) = B_1 + B_2$. Since the theorem is proved for completely regular spaces there is a decomposition

$$\rho(X) = Y_1 \cup Y_2$$

into disjoint closed sets such that $B_1 = B(Y_1)$ and $B_2 = B(Y_2)$. From the definition of $\rho(X)$ it is clear that there is a decomposition

$$X = X_1 \cup X_2$$

into disjoint closed sets such that $Y_1 = \rho(X_1)$ and $Y_2 = \rho(X_2)$. Consequently $B_1 = B(X_1)$ and $B_2 = B(X_2)$.

From Theorem 9.2 we obtain the following corollaries:

COROLLARY 9.3. *X is connected if and only if $B(X)$ is indecomposable as a direct sum.*

COROLLARY 9.4. *X contains at least n components if and only if $B(X)$ is a direct sum of n summands.*

COROLLARY 9.5. *X contains an isolated point if and only if $B(X)$ admits the straight line as a direct summand.*

UNIVERSITY OF MICHIGAN.

SIMPLIZIALZERLEGUNGEN VON BESCHRÄNKTER FLACHHEIT¹

BY HANS FREUDENTHAL

(Received July 25, 1941)

Wir beantworten eine Frage von Herrn Prof. L. E. J. Brouwer nach einer einfachen Konstruktion einer unendlichen Folge beliebig feiner Simplizialzerlegungen eines Polytops, deren jede eine Unterteilung der vorangehenden ist, und bei der die auftretenden Teilsimplexe nicht beliebig flach werden dürfen, d. h. bei der der Quotient

$$c^r/v$$

(c = Durchmesser = größte Kantenlänge des Simplexes, v = Volumen des Simplexes) für alle r -dimensionalen Simplexe gleichmäßig beschränkt bleibt. Derartige Zerlegungsfolgen braucht man in der Analysis und auch im Grenzgebiet zwischen Analysis und Topologie.

Die Konstruktion, die wir hier angeben wollen, verläuft ähnlich wie unsere Simplizialzerlegung des Cartesischen Produkts zweier Simplexe.²

1

Simplexe sollen hier immer mit einer *festen* Reihenfolge der Ecken gegeben sein, Parallelotope immer unter Auszeichnung einer Ecke und einer Anordnung der von dieser Ecke ausgehenden Kanten.

Das Simplex T habe die Ecken

$$e_0, \dots, e_r.$$

Wir können T auch beschreiben durch die Vektoren

$$x_0 = e_0, \quad x_i = e_i - e_{i-1} \quad (i = 1, \dots, r).$$

Sei π eine Permutation p_1, \dots, p_r der Zahlen $1, \dots, r$. Die Punkte

$$e_i = x_0 + \sum_{p=1}^r x_{p_i} \quad (i = 0, \dots, r)$$

(als Ecken) erzeugen ein Simplex T^π . Ist π_0 die identische Permutation der Zahlen $1, \dots, r$, so ist $T^{\pi_0} = T$. Die $r!$ Simplexe T^π heißen *untereinander konjugiert*.

T^π ist die Gesamtheit der Punkte

$$T^\pi \dots \sum_{p=1}^r \lambda_p e_{p_i} \quad \text{mit} \quad \lambda_p \geq 0, \quad \sum_{p=1}^r \lambda_p = 1,$$

¹ Diese Note stimmt im Wesentlichen überein mit einer im März 1939 bei den *Fundamenta Mathematicae* eingereichten Note die dort nicht mehr erschienen ist. Das Ergebnis habe ich bereits anderswo verwendet [Proceedings Amsterdam 42 (1939), 880-901].

² Fund. Math. 29 (1937), 138-144.

oder, was auf dasselbe hinauskommt,

$$T^\pi \cdots x_0 + \sum_{r=1}^r \alpha_r x_{p_r} \quad \text{mit} \quad 1 \geq \alpha_1 \geq \cdots \alpha_r \geq 0.$$

Daraus folgt: Die T^π bilden eine simpliziale Teilung des Parallelotops P ,

$$P \cdots x_0 + \sum \alpha_r x_r, \quad 0 \leq \alpha_r \leq 1.$$

Umgekehrt wird durch den Parallelotop P , von dem die Ecke x_0 und die Kantenvektoren x_1, \dots, x_r gegeben sind, und durch die Permutation π eindeutig das Simplex T^π bestimmt.

2

Aus dem Parallelotop P entstehen durch Halbierung quer zu allen Kanten die Parallelotope

$$P_\sigma \cdots x_{\sigma,0} + \frac{1}{2} \sum \alpha_r x_r;$$

hier sei σ ein System

$$s_1, \dots, s_r$$

von Zahlen s_r ,

$$s_r = 0 \text{ oder } 1,$$

und

$$x_{\sigma,0} = x_0 + \frac{1}{2} \sum s_r x_r.$$

In derselben Weise wie P in die Simplexe T^π ist jedes P_σ zerlegt in die Simplexe $(T_\sigma)^\pi$.— T^π und $(T_\sigma)^\pi$ sind homothetisch (1:2).

Sei π wieder die Permutation p_1, \dots, p_r der $1, \dots, r$ und sei $\pi^* = \pi^\sigma$ die Permutation p_1^*, \dots, p_r^* , die entsteht, wenn man erst (in der durch π gegebenen Reihenfolge) alle p_r mit $s_{p_r} = 1$ auszieht, und dann alle mit $s_{p_r} = 0$, also:

für $s_a > s_b$ kommt a vor b in π^* ,

für $s_a = s_b$ und a vor b in π kommt auch a vor b in π^* .

Wir behaupten: $(T_\sigma)^\pi$ liegt ganz in T^{π^*} .

In der Tat ist

$$\begin{aligned} (T_\sigma)^\pi \cdots x_0 + \frac{1}{2} \sum_{r=1}^r s_r x_r + \frac{1}{2} \sum_{r=1}^r \alpha_r x_{p_r} \\ = x_0 + \frac{1}{2} \sum' (\alpha_r + \frac{1}{2}) x_{p_r} + \frac{1}{2} \sum'' \alpha_r x_{p_r} \quad (1 \geq \alpha_1 \geq \cdots \geq \alpha_r \geq 0) \\ (\text{wo } \sum' \text{ zu erstrecken ist über alle } r \text{ mit } s_{p_r} = 1 \text{ und } \sum'' \text{ über alle } r \text{ mit } s_{p_r} = 0) \\ = x_0 + \sum_{r=1}^r \beta_r x_{p_r}; \quad (1 \geq \beta_1 \geq \cdots \geq \beta_u \geq \frac{1}{2} \geq \beta_{u+1} \geq \cdots \geq \beta_r \geq 0) \end{aligned}$$

(wo u die Zahl der Einsen unter den s_r ist). Hieraus folgt unsere Behauptung.

Weiter folgt hieraus: Die $(T_\sigma)^\pi$ mit $\pi^\sigma = \pi^*$ bilden eine Simplizialzerlegung $Z(T)$ von T .

3

Sei e_{ij}^* der Mittelpunkt der Strecke $e_i^* e_j^*$, $e_{ii}^* = e_i^*$,

$$\begin{aligned} e_{ij}^* &= \frac{1}{2} \left(x_0 + \sum_{\nu=1}^i x_{\nu} + x_0 + \sum_{\nu=1}^j x_{\nu} \right) \\ &= x_0 + \sum_{\nu=1}^i x_{\nu} + \frac{1}{2} \sum_{\nu=i+1}^j x_{\nu} \quad (i \leq j). \end{aligned}$$

Seien unter den s_ν , mit $\nu \leq k$

k' Einsen und k'' Nullen.

Die k^{te} Ecke von $(T_r)^*$ erhält man für $\alpha_1 = \dots = \alpha_k = 1$, $\alpha_{k+1} = \dots = \alpha_r = 0$; sie lautet also

$$x_0 + \sum_{\nu=1}^{k'} x_{\nu} + \frac{1}{2} \sum_{\nu=i'+1}^{i'+u} x_{\nu} = e_{ij}^* \quad (i = k', j = k'' + u).$$

Daraus folgt für jedes Teilsimplex $(T_r)^*$ von T^{**} :

0^{te} Ecke ist e_{0u}^ ;*

auf die Ecke e_{ij}^ folgt die Ecke $e_{i+1,j}^*$ oder die Ecke $e_{i,i+1}^*$;*

der erste Index ist stets $\leq u$, der zweite $\leq r$.

Umgekehrt sind die Teilsimplexe $(T_r)^*$ von T^{**} durch diese Eigenschaften auch gekennzeichnet.

4

Wir achten nun nur noch auf $T = T^{*0}$ und lassen den Index π_0 im Folgenden weg. Wir können dann die Teilung $Z(T)$ von T auch so beschreiben:

Die Ecken von $Z(T)$ sind die e_{ij} ;

dann und nur dann bilden e_{ij} und $e_{i'j'}$ ein eindimensionales Simplex, wenn die Paare ij und $i'j'$ einander nicht trennen;

eine Menge von e_{ij} bildet dann und nur dann ein Simplex von $Z(T)$, wenn ihre Elemente zu je zweien Simplexe von $Z(T)$ bilden;

in den Simplexen von $Z(T)$ sind die Ecken nach steigenden $i + j$ angeordnet.

Wir bemerken weiter:

Die Simplexe von $Z(T)$ sind ähnlich der zu T konjugierten.

Sei nun ein endlicher Polytop R gegeben! Wir legen eine Reihenfolge der Ecken zugrunde und bilden die Teilung $Z(R)$, indem wir auf jedes Simplex von R den Prozeß Z anwenden. Die Ecken von $Z(R)$ ordnen wir lexikographisch. Indem wir den Prozeß Z unbegrenzt wiederholen, erhalten wir eine Unterteilungsfolge. Alle dabei auftretenden Simplexe sind den (endlich vielen) konjugierten der Simplexe von R ähnlich, und da ähnliche Simplexe dieselbe Flachheit

$$c'/v$$

besitzen, bleibt, wie wir es wünschten, die Flachheit gleichmäßig beschränkt.

A SIMPLIFIED PROOF FOR THE RESOLUTION OF SINGULARITIES OF AN ALGEBRAIC SURFACE

BY OSCAR ZARISKI

(Received March 19, 1942)

1. Introduction

We presuppose the theorem of local uniformization, as proved elsewhere.¹ By this theorem, any valuation of a field Σ/\mathbf{K} of algebraic functions of r independent variables, over a given ground field \mathbf{K} of characteristic zero, can be "uniformized" on a suitable projective model V of Σ , i.e. the center of the valuation on a suitable model V will be a simple subvariety of V .² If the ground field \mathbf{K} (the field of coefficients) is the field of complex numbers (the classical case), then the above result, in conjunction with the bicomcompactness of the Riemann manifold \mathbf{M} of Σ/\mathbf{K} ,³ implies the existence of a *finite* set of models of Σ , say

$$V_1, V_2, \dots, V_n,$$

such that any zero-dimensional valuation of Σ/\mathbf{K} is uniformized on at least one of the models V_i of the set.⁴ Any finite set of projective models with the above property shall be called a *resolving system* of the Riemann manifold \mathbf{M} .

In the case of abstract varieties, where we cannot fall back on topology, it is necessary to give an algebraic proof of the existence of resolving systems of \mathbf{M} . In the special case of algebraic surfaces the algebraic proof of the existence of resolving systems is strikingly simple (see section 6 of this paper). The general case of algebraic varieties will be treated in a subsequent paper.

The theorem of the resolution of singularities of an algebraic variety can be formulated in terms of resolving systems, as follows: *There exists a resolving system of the Riemann manifold \mathbf{M} which consists of only one model.* In view of the existence of resolving systems, this theorem will be established if it can be

¹ For the case of algebraic surfaces, see our paper "The reduction of singularities of an algebraic surface", *Annals of Mathematics*, vol. 40 (July, 1939), pp. 649-659. For the general case of varieties of any dimension, see our paper "Local uniformization on algebraic varieties", *Annals of Mathematics*, vol. 41 (October, 1940), pp. 852-896. These two papers will be referred to respectively as "Reduction" and "Uniformization".

² For the definition of the center of a valuation see "Uniformization", p. 857.

³ \mathbf{M} is the totality of all *zero-dimensional* valuations (or *places*) of Σ/\mathbf{K} ; see "Uniformization", p. 855.

⁴ It follows then necessarily that *any* valuation (whether of dimension zero or greater than zero) will be uniformized on at least one of the models V_i . To see this, it is only necessary to observe that: a) If B is any valuation of dimension > 0 , then there exist zero-dimensional valuations compounded with B ; b) if B_0 is such a zero-dimensional valuation then on every model V of Σ the center of B_0 will lie on the center of B ; c) if an irreducible subvariety W of V contains a simple point, then W itself is simple.

proved that *the existence of a resolving system of \mathbf{M} consisting of n models ($n > 1$) implies the existence of a resolving system of \mathbf{M} consisting of $n - 1$ models.* To carry out this induction from n to $n - 1$, it is sufficient to prove the *fundamental theorem* which we are now going to state.

Let N be an arbitrary subset of \mathbf{M} , i.e. let N be an arbitrary set of places of our field Σ/\mathbf{K} . In the same fashion as we have defined resolving system of \mathbf{M} , we can define resolving systems of N . A resolving system of N will be therefore any finite set of models of Σ such that any valuation in N has a simple center on at least one of the models in the set. In particular, if a resolving system of N consists of only one model, this model shall be called a *resolving model* for N .

FUNDAMENTAL THEOREM. *If N is an arbitrary subset of \mathbf{M} and if there exists a resolving system of N consisting of two models, then there also exists a resolving model for N .*

From the fundamental theorem, the theorem on the resolution of singularities follows immediately. For, let V_1, \dots, V_n be a resolving system of \mathbf{M} . Let N be the subset of \mathbf{M} consisting of those valuations which have a singular center on *each* of the models V_1, \dots, V_{n-2} . It is clear that the pair V_{n-1}, V_n constitutes a resolving system for N . If we assume the fundamental theorem, then there exists a resolving model V'_{n-1} for N . The $n - 1$ models $V_1, \dots, V_{n-2}, V'_{n-1}$ constitute a resolving system for \mathbf{M} , and our induction from n to $n - 1$ is complete.

The aim of this paper is to give a proof of the fundamental theorem in the case of algebraic surfaces. Our present proof for the resolution of singularities is much simpler than our earlier proof (see "Reduction") and is also more general, since at present we do not assume that the ground field \mathbf{K} is algebraically closed. In the course of the proof we have to make use of certain properties of fundamental loci of birational correspondences. This preliminary material, strictly confined to the precise needs of our proof, is presented in the next section. In a forthcoming paper dealing with the general theory of birational correspondences we study these properties systematically. A brief account of this general theory will be found in my address "Normal varieties and birational correspondences" (delivered before the annual meeting of the Society in Bethlehem in 1941) which has appeared in the Bulletin of the American Mathematical Society, Vol. 48, No. 6 (June 1942).

I. PRELIMINARY CONCEPTS

2. Birational correspondences

Let V and V' be two models of our field Σ/\mathbf{K} .

DEFINITION. Two points, P, P' of V and V' respectively are corresponding points if there exists a valuation of Σ/\mathbf{K} such that its center on V is the point P and its center on V' is the point P' .⁵

⁵ Compare with p. 666 of "Reduction". Our study of the general theory of birational correspondences is based on this valuation-theoretic definition.

We say that P is a *fundamental point* of the birational correspondence between V and V' , if there exists a corresponding point P' on V' such that $Q(P) \not\subseteq Q(P')$ (here $Q(P)$ denotes the quotient ring of P).

Suppose that there exists a point P' which corresponds to P and which is such that $Q(P') \subseteq Q(P)$. If v is a valuation of centers P and P' on V and V' respectively and if \mathfrak{P}_v denotes the prime ideal of v , then $\mathfrak{P}_v \cap Q(P) = \mathfrak{p}$ and $\mathfrak{P}_v \cap Q(P') = \mathfrak{p}'$, where $\mathfrak{p}(\mathfrak{p}')$ is the ideal of non units in $Q(P)(Q(P'))$. Hence $\mathfrak{p} \cap Q(P') = \mathfrak{p}'$, and from this it follows that *any* valuation whose center on V is P has P' as center on V' , i.e. P' is the only point of V' which corresponds to P . Therefore, if P is a point of V which is not fundamental, then to P there corresponds a unique point P' on V' , and we have $Q(P') \subseteq Q(P)$. On the other hand, if P is fundamental, then it follows that $Q(P) \not\subseteq Q(P')$ for any point P' which corresponds to P .

In the case of surfaces we have proved elsewhere that if V and V' are *normal* surfaces and if to P there corresponds a finite number of points on V' , then P is not fundamental.⁶ Therefore, if P is fundamental, the locus of corresponding points P' is a variety of dimension 1.⁷

If neither V nor V' carry fundamental points, then the birational correspondence between V and V' shall be called *regular*.

In any case, the number of fundamental points on either surface is always finite.

Let $\xi_0^*, \xi_1^*, \dots, \xi_n^*$ be the homogeneous coördinates of the general point of V , and let $\eta_0^*, \eta_1^*, \dots, \eta_m^*$ be the homogeneous coördinates of the general point⁸ of V' . The $(n+1)(m+1)$ products

$$\omega_{ij}^* = \xi_i^* \eta_j^*$$

can be regarded as homogeneous coördinates of the general point of a variety V^* which is birationally equivalent to V and to V' : for the quotient of any two ω^* 's, say $\omega_{ij}^*/\omega_{\alpha\beta}^*$, is equal to $\xi_i^*/\xi_\alpha^* \cdot \eta_j^*/\eta_\beta^*$, and is therefore an element of the field Σ . Moreover, the non-homogeneous coördinates of the general point of V^* , i.e. the quotients of the ω^* 's by a fixed ω^* , say by ω_{00}^* , generate the field Σ/K , since the quotients ξ_i^*/ξ_0^* (and also the quotients η_j^*/η_0^*) are among them. The variety V^* is called the *join* of V and V' . It is clear that the ring

⁶ See "Reduction", p. 688, Theorem 5, but the proof is much simpler and is as follows. If P'_1, \dots, P'_h are the points on V' which correspond to P , then for any valuation v whose center is P it must be true that the valuation ring of v contains one of h quotient rings $Q(P'_i)$, whence it also contains the intersection \mathfrak{J} of these quotient rings. Since $Q(P)$ is integrally closed, it is the intersection of the above mentioned valuation rings. Hence $Q(P) \supseteq \mathfrak{J}$. If $\mathfrak{p} \cap \mathfrak{J} = \mathfrak{p}$, then \mathfrak{p} is one of the h prime maximal ideals of \mathfrak{J} , say the one relative to the point P'_1 , and it is clear that any valuation of center P will have center P'_1 on V' , i.e. $h = 1$. [For the definition of normal varieties, see our paper "Some results in the arithmetic theory of algebraic varieties", p. 282, American Journal of Mathematics, vol. 61 (April 1939). This paper will be referred to as "Results".]

⁷ See "Reduction", p. 667.

⁸ For the concept of homogeneous coördinates of the general point of an irreducible algebraic variety, see "Results" p. 284.

$K[\omega_{10}^*/\omega_{00}^*, \dots, \omega_{nm}^*/\omega_{00}^*]$ of polynomials in these non-homogeneous coordinates—briefly: the ring of non-homogeneous coordinates—coincides with the ring $K[\xi_1^*/\xi_0^*, \dots, \xi_n^*/\xi_0^*, \eta_1^*/\eta_0^*, \dots, \eta_m^*/\eta_0^*]$ and is therefore the join of the corresponding rings of non-homogeneous coordinates of the general points of V and V' . From this one concludes immediately that the birational correspondence between V^* and either one of the two given models V, V' has no fundamental points on V^* .

It follows that each point P^* of V^* represents a unique pair (P, P') of corresponding points of V and V' . Conversely, each such pair (P, P') is represented by at least one point P^* of the join V^* . For this reason the join V^* is often referred to in the literature as the variety of pairs of corresponding points of V and V' . However, this last term may be misleading when we deal with a ground field K which is not algebraically closed. For in this case (and only in this case) it may happen that one and the same pair (P, P') of corresponding points of V and V' is represented by (or, we may say, splits into) more than one point of V^* .

The join of any finite number of models V_1, \dots, V_h can be defined as the join of V_h and of the join of V_1, \dots, V_{h-1} . It is seen immediately that this definition is symmetric in V_1, \dots, V_h .

3. Quadratic transformations

Let P be a point of V and let \mathfrak{p}^* be the corresponding prime homogeneous ideal (of projective dimension zero) in the ring $K[\xi_0^*, \xi_1^*, \dots, \xi_n^*]$. Let $\eta_0^*, \eta_1^*, \dots, \eta_m^*$ be a set of forms in $\xi_0^*, \xi_1^*, \dots, \xi_n^*$, of like degree, such that the ideal $(\eta_0^*, \eta_1^*, \dots, \eta_m^*)$ differs from \mathfrak{p}^* only by an irrelevant primary component i.e. a component belonging to the irrelevant prime ideal $\mathfrak{p}_0 = (\xi_0^*, \xi_1^*, \dots, \xi_n^*)$.⁹ By a *quadratic transformation of center P* we mean the birational transformation which carries V into the variety \bar{V} whose general point has the following homogeneous coordinates:

$$(1) \quad \omega_{ij}^* = \xi_i^* \eta_j^*, \quad i = 0, 1, \dots, n; j = 0, 1, \dots, m.$$

This transformation depends of course on the choice of the forms η_j^* , but in a non-essential fashion. Namely, if \bar{V}_1 is the transform of V by a quadratic transformation, relative to another set of forms, then \bar{V} and \bar{V}_1 are in regular birational correspondence.

PROOF. Let $\xi_0^*, \xi_1^*, \dots, \xi_\mu^*$ be the set of forms which defines \bar{V}_1 , and let $\omega_{ij}^* = \xi_i^* \xi_j^*$. Let: $\alpha = \text{degree } \eta_j^*, \beta = \text{degree } \xi_j^*, \mathfrak{A} = (\eta_0^*, \dots, \eta_m^*), \mathfrak{B} = (\xi_0^*, \dots, \xi_\mu^*)$. Since the ideals \mathfrak{A} and \mathfrak{B} differ only in their primary irrelevant components, we will have for sufficiently high integers a and b : $\mathfrak{A} \cdot \mathfrak{p}_0^a \equiv 0(\mathfrak{B}), \mathfrak{B} \cdot \mathfrak{p}_0^b \equiv 0(\mathfrak{A})$. Select a and b so as to have $\alpha + a = \beta + b$.

⁹ Let $\mathfrak{p}_0^* = (\phi_0, \phi_1, \dots, \phi_s)$, where the ϕ_i are forms in the ξ_j^* 's and let $\nu_i = \text{degree of } \phi_i, \nu = \max. (\nu_0, \nu_1, \dots, \nu_s)$. Then the forms $\xi_j^{*-\nu_i} \phi_i, j = 0, 1, \dots, n, i = 0, 1, \dots, s$, satisfy the desired condition.

The elements ω_{ij}^* constitute a linear base for the forms of degree $\alpha + 1$ which belong to \mathfrak{A} . If we multiply the ω_{ij}^* by ξ_0^*, \dots, ξ_n^* , then the products give a base for the forms of degree $\alpha + 2$ in \mathfrak{A} . These products are the homogeneous coördinates of the general point of the join V^* of V and \bar{V} . Since, as will be pointed out later on, the quadratic transformation has no fundamental points on \bar{V} , it follows that \bar{V} and V^* are in regular birational correspondence. We now multiply the homogeneous coördinates of the general point of V^* by ξ_0^*, \dots, ξ_n^* , getting the join of V^* and V . Proceeding in this fashion we construct a model V^* , such that: a) the homogeneous coördinates of the general point of V^* constitute a linear base for the forms of degree $\alpha + a$ in \mathfrak{A} ; i.e., they are elements of $\mathfrak{A}p_0^a$, and consequently also of \mathfrak{B} ; b) V^* and \bar{V} are in regular birational correspondence. In a similar fashion we construct a model V_1^* in regard to \mathfrak{B} , \bar{V}_1 and the integer $\beta + b$. Since $\alpha + a = \beta + b$, it follows immediately that the homogeneous coördinates of the general point of V^* are linearly dependent on the homogeneous coördinates of the general point of V_1^* , and *vice versa*. Consequently, V^* and V_1^* are related projectively to each other, q.e.d.

If K is algebraically closed, then it is permissible to assume that the coördinates of P (elements of K) are $1, 0, 0, \dots, 0$. Then $\mathfrak{p}^* = (\xi_1^*, \dots, \xi_n^*)$, and our quadratic transformation is given by the equations: $\omega_{ij} = \xi_i \xi_j$, $i = 0, 1, \dots, n$; $j = 1, 2, \dots, n$. This is the ordinary quadratic transformation defined by the system of hyperquadrics passing through the point P (see "Reduction," p. 676).

Since P is the only point of V at which all the forms $\xi_i^* \eta_j^*$ vanish simultaneously, P is the only fundamental point of the quadratic transformation. To any other point A of V there corresponds a unique point \bar{A} on \bar{V} , and moreover $Q(\bar{A})$ coincides with $Q(A)$. The transformation has no fundamental points on \bar{V} . The proofs of all these assertions are straightforward and do not differ essentially from the proofs in the case of an algebraically closed ground field (see "Reduction," p. 676–679).

To see what happens to the point P , we pass to non-homogeneous coördinates: $\xi_i = \xi_i^* / \xi_0^*$ ($i \neq 0$), and $\omega_{ij} = \omega_{ij}^* / \omega_{00}^*$ (i, j not both zero). Let η_i denote the non-homogeneous polynomial obtained from the form η_i^* by the usual process (replace ξ_0^* by 1 and ξ_i^* by ξ_i , $i \neq 0$). The ring of non-homogeneous coördinates for V , resp. \bar{V} , will be: $\mathfrak{o} = K[\xi_1, \dots, \xi_n]$ and

$$\bar{\mathfrak{o}} = K[\xi_1, \dots, \xi_n, \eta_1/\eta_0, \dots, \eta_m/\eta_0]$$

respectively. The point P will be given by the prime ideal $\mathfrak{p} = (\eta_0, \dots, \eta_m)$ in \mathfrak{o} . Since the ideal $\bar{\mathfrak{o}}\mathfrak{p}$ is the principal ideal $\bar{\mathfrak{o}} \cdot \eta_0$, we conclude immediately that the transform $T(P)$ of the point P is a pure $(r - 1)$ -dimensional subvariety of \bar{V} (not necessarily irreducible).¹⁰

¹⁰ If Γ is an irreducible subvariety of \bar{V} , at finite distance with respect to the non-homogeneous coördinates ω_{ij} , Γ is given by a prime ideal $\bar{\mathfrak{p}}$ in the ring $\bar{\mathfrak{o}}$. If Γ corresponds to the

4. Quadratic transformations with simple center. The p -adic divisor

We are now especially interested in the case in which P is a simple point. Let \mathfrak{J} be the quotient ring of P and let \mathfrak{P} denote the ideal of non-units in \mathfrak{J} . Let $t_1, \dots, t_r, t_i \in \mathfrak{J}$, be uniformizing parameters of P , i.e. a set of r elements in \mathfrak{J} with the property $\mathfrak{J} \cdot (t_1, \dots, t_r) = \mathfrak{P}$.¹¹

A significant property of the uniformizing parameters is the following: if $\phi(t_1, \dots, t_r) = 0$, where ϕ is a polynomial in t_1, \dots, t_r with coefficients in \mathfrak{J} , and if $\phi_p(t_1, \dots, t_r)$ is the sum of terms of ϕ of lowest degree ρ (ϕ_p = the *leading form* of ϕ), then the coefficients of ϕ_p must all be divisible by \mathfrak{P} .¹² This property enables us to construct a valuation of Σ in the following fashion. If $\alpha \in \mathfrak{J}$ and if α is exactly divisible by \mathfrak{P}^ρ , then $\alpha = \phi_p(t_1, \dots, t_r)$, where ϕ_p is a form of degree ρ with coefficients in \mathfrak{J} but not all in \mathfrak{P} . Let β be another element of \mathfrak{J} , exactly divisible by \mathfrak{P}^σ , so that $\beta = \psi_\sigma(t_1, \dots, t_r)$. We have $\alpha\beta \equiv 0(\mathfrak{P}^{\rho+\sigma})$, and since the coefficients of the form $\phi_p\psi_\sigma$ are not all in \mathfrak{P} , we conclude, by the property of the uniformizing parameters stated above, that $\alpha\beta$ is not divisible by $\mathfrak{P}^{\rho+\sigma+1}$. This shows that if we put $v(\alpha) = \rho$, we get a *discrete valuation* B of Σ , of rank 1. We shall call B the *p -adic divisor of center P* (P — a simple point!).

It is not difficult to see that B is of dimension $r - 1$, i.e. B is a divisor. For, we have $v(t_i) = 1$, $v(t_i/t_1) = 0$. Were the B -residues of $t_2/t_1, \dots, t_r/t_1$ algebraically dependent over K , there would exist a form $\phi_p(t_1, \dots, t_m)$, with coefficients in K , not all zero, such that $v(\phi_p) > v(t_1^m)$. This would imply $\phi_p \equiv 0(\mathfrak{P}^{\rho+1})$, which is impossible.

The following theorem puts into evidence the effect of a quadratic transformation of center P in regard to our p -adic divisor:

THEOREM 1. *If \bar{V} is the transform of V under a quadratic transformation T of simple center P , then the p -adic divisor of center P is of the first kind with respect to \bar{V} , i.e. the center of the divisor on \bar{V} is $(r - 1)$ -dimensional. This center coincides with the transform $T(P)$ (whence $T(P)$ is irreducible).*

PROOF. We use the notations of the preceding section. We assume that the variety $T(P)$ is not entirely at infinity, i.e. that the principal ideal $\mathfrak{d} \cdot \eta_0$ is not unit ideal (see footnote 10).

point P , then we must have $\mathfrak{p} \cap \mathfrak{o} = \mathfrak{p}$ (see footnote 7). Hence $\mathfrak{d}\mathfrak{p} = \mathfrak{o}(\mathfrak{p})$, i.e. Γ lies on the subvariety \bar{W} of \bar{V} defined by the principal ideal $\mathfrak{d} \cdot \eta_0$. Conversely, if Γ is on \bar{W} , then $\mathfrak{d} \cdot \mathfrak{p} \equiv 0(\mathfrak{p})$, whence $\mathfrak{p} \cap \mathfrak{o} = 0$ (since \mathfrak{p} is a maximal ideal of \mathfrak{o}), and therefore Γ corresponds to P . This shows that $T(P)$ consists of \bar{W} and perhaps of other irreducible components at infinity. Since any irreducible component of $T(P)$ can be assumed to be at finite distance, for a proper choice of the non-homogeneous coördinates (if $\omega_{\alpha\beta}^*$ is different from zero on the given component, we use the quotients $\omega_{\alpha\beta}^*/\omega_{\alpha\beta}^*$) and since if the principal ideal $\mathfrak{d} \cdot \eta_0$ is not the unit ideal, all its isolated prime ideals are $(r - 1)$ -dimensional, our assertion follows.

¹¹ See our paper "Algebraic varieties over ground fields of characteristic zero", American Journal of Mathematics, vol. 62 (January, 1940), p. 199.

¹² See loc. cit. in footnote 11, p. 202 and p. 207-208. As a consequence of this property, the quotient ring of a simple point (and also, more generally, the quotient ring of a simple subvariety) is a " p -Reihenring" in the sense of Krull (see Krull, Dimensionstheorie in Stellenringen, Crelle's Journal, vol. 179 (1938)).

To prove the theorem it will be sufficient to prove that the ideal $\mathfrak{F} \cdot \eta_0$ is prime, where $\mathfrak{F} = \mathfrak{F} \cdot \bar{\mathfrak{o}}$ (\mathfrak{F} = quotient ring of P),¹³ and that the irreducible subvariety of \bar{V} defined by this prime ideal is the center of B . Let B' be an arbitrary valuation of center P , whose center on \bar{V} is at finite distance (such valuations exist, since we have assumed that $T(P)$ is not entirely at infinity).

Among the uniformizing parameters t_1, \dots, t_r of P we take one which has least value in B' . Let it be t_1 . Since $\mathfrak{P} = \mathfrak{F} \cdot (\eta_0, \eta_1, \dots, \eta_m)$, it follows that $t_i \in \mathfrak{F} \cdot (\eta_0, \eta_1, \dots, \eta_m)$. Consequently $t_i/\eta_0 \in \mathfrak{F}$, and, in particular, $t_1/\eta_0 \in \mathfrak{F}$. Since the center of B' is at finite distance we conclude that $v_{B'}(t_1) \geq v_{B'}(\eta_0)$. This inequality implies that $\eta_0 \not\equiv 0(\mathfrak{P}^2)$. For, assume that $\eta_0 \equiv 0(\mathfrak{P}^2)$. Then $\eta_0 = \phi_2(t_1, \dots, t_r)$, where ϕ_2 is a quadratic form in the t 's, with coefficients in \mathfrak{F} . We can write $\eta_0/t_1 = t_1\phi_2(1, t_2/t_1, \dots, t_r/t_1)$. By hypothesis, $v_{B'}(t_i/t_1) \geq 0$. We also have $v_{B'}(t_1) > 0$ and $v_{B'}(\alpha) \geq 0$, for any element α in \mathfrak{F} (since P is the center of B'). Consequently, $v_{B'}(\eta_0/t_1) > 0$, a contradiction.

Since $\eta_0 \not\equiv 0(\mathfrak{P}^2)$, we have $v_B(\eta_0) = 1$, and since $v_B(\eta_i) \geq 1$, we conclude that the quotients η_i/η_0 belong to the valuation ring of B . Consequently the entire ring \mathfrak{F} is contained in the valuation ring of B . Therefore the center of B on V is a (irreducible) subvariety \bar{W} of V at finite distance, where \bar{W} will be given by the prime ideal $\bar{\mathfrak{P}}$ in \mathfrak{F} consisting of the elements of positive value in B . Now it is clear that $\mathfrak{F} \cdot \eta_0 \equiv 0(\bar{\mathfrak{P}})$. On the other hand, let $\bar{\alpha} \in \mathfrak{F}$, $v(\bar{\alpha}) > 0$. We can write $\bar{\alpha}$ in the form: $\bar{\alpha} = \phi_\rho(\eta_0, \eta_1, \dots, \eta_m)/\eta_0^\rho$, where ϕ_ρ is a form of degree ρ , with coefficients in \mathfrak{F} . Since $v_B(\bar{\alpha}) > 0$ and $v_B(\eta_0^\rho) = \rho$, we must have $v_B(\phi_\rho) \geq \rho + 1$. Hence $\phi_\rho = \psi_{\rho+1}(t_1, \dots, t_r)$, where $\psi_{\rho+1}$ is a form of degree $\rho + 1$ in t_1, \dots, t_r , with coefficients in \mathfrak{F} . Since $t_i \equiv 0(\eta_0, \eta_1, \dots, \eta_m)$, we will also have $\phi_\rho = g_{\rho+1}(\eta_0, \dots, \eta_m)$, where $g_{\rho+1}$ is again a form of degree $\rho + 1$ in η_0, \dots, η_m , with coefficients in \mathfrak{F} . Consequently, $\bar{\alpha} = g_{\rho+1}/\eta_0^\rho = \eta_0 \cdot g_{\rho+1}(1, \eta_1/\eta_0, \dots, \eta_m/\eta_0)$, whence $\bar{\alpha} \equiv 0(\mathfrak{F} \cdot \eta_0)$. We conclude that $\mathfrak{F} \cdot \eta_0 = \bar{\mathfrak{P}}$, and this completes the proof.

From the fact that the irreducible variety $T(P)$ is defined by a principal ideal, namely by the ideal $\bar{\mathfrak{o}} \cdot \eta_0$, it follows $T(P)$ is a simple subvariety of \bar{V} (of dimension $r - 1$, and having η_0 as uniformizing parameter). Therefore $T(P)$ contains points which are simple for \bar{V} . We shall need, however, the following stronger result:

THEOREM 2. Every point of $T(P)$ is a simple point of \bar{V} .

PROOF. Let \bar{P} be a point of $T(P)$, which we may assume to be a point at finite distance. We consider an arbitrary but fixed valuation B' with centers P and \bar{P} (on V and \bar{V} respectively). Assuming, as we did before, that $v_{B'}(t_i) \leq v_{B'}(t_i)$, $i = 1, 2, \dots, r$, we will have $t_i/\eta_0 \in \mathfrak{F}$, and since $\mathfrak{F} \subset Q(\bar{P})$, it follows that

$$(2) \quad \frac{t_i}{\eta_0} \in Q(\bar{P}), \quad i = 1, 2, \dots, r.$$

¹³ Every prime ideal $\bar{\mathfrak{p}}$ of $\bar{\mathfrak{o}} \cdot \eta_0$ contracts to \mathfrak{p} (see footnote 11), and hence $\mathfrak{F} = \mathfrak{o}_P \subseteq \bar{\mathfrak{o}}_P$. From this it follows immediately the ideals $\bar{\mathfrak{o}} \cdot \eta_0$ and $\mathfrak{F} \cdot \eta_0$ have like decompositions into maximal primary components.

We consider the ring $\mathfrak{o}^* = K[\xi_1, \dots, \xi_n, t_2/t_1, \dots, t_r/t_1]$. Let V^* be the model whose general point (in non-homogeneous coördinates) is $(\xi_1, \dots, \xi_n, t_2/t_1, \dots, t_r/t_1)$. Since $v_{B'}(t_i/t_1) \geq 0$, the center P^* of B' on V^* is a point at finite distance. We will have also the following relations, similar to (2):

$$(3) \quad \frac{\eta_i}{t_1} \in Q(P^*), \quad i = 0, 1, \dots, m.$$

From (2) and (3) it follows that t_1/η_0 is an element of $Q(\bar{P})$ and that its reciprocal has non-negative value in the valuation B' . Since \bar{P} is the center of B' on \bar{V} , we conclude that t_1/η_0 is a unit in $Q(\bar{P})$. Similarly, t_1/η_0 is a unit in $Q(P^*)$. But then the quotients $t_i/t_1 (= t_i/\eta_0 \cdot \eta_0/t_1)$ are in $Q(\bar{P})$, and the quotients $\eta_i/\eta_0 (= \eta_i/t_1 \cdot t_1/\eta_0)$ are in $Q(P^*)$. Therefore $\bar{\mathfrak{o}} \subset Q(P^*)$ and $\mathfrak{o}^* \subset Q(\bar{P})$, and this implies: $Q(P^*) = Q(\bar{P})$. Thus, we have only to show that P^* is a simple point of V^* . For that we have to exhibit uniformizing parameters at P^* .

Let Δ be the residue class field of the point P , i.e. let $\Delta = \mathfrak{o}/\mathfrak{p}$. Here Δ is an algebraic extension of K . Similarly, let Δ^* be the residue class field of P^* . We have $\Delta^* \supseteq \Delta$, since $\mathfrak{o}^* \supseteq \mathfrak{o}$ and since the prime \mathfrak{o}^* -ideal \mathfrak{p}^* of the point P^* contracts in \mathfrak{o} to \mathfrak{p} . Let c_1, \dots, c_n be the P -residues of ξ_1, \dots, ξ_n ($c_i \in \Delta$), and let d_2, \dots, d_r be the P^* -residues of $t_2/t_1, \dots, t_r/t_1$ ($d_i \in \Delta^*$). The residue d_i will be the root of an irreducible polynomial $f_i(z)$ with coefficients in Δ . Replace each coefficient of $f_i(z)$ by an arbitrary but fixed element of \mathfrak{o} of which it is the residue. We get a certain polynomial $F_i(z)$ with coefficients in \mathfrak{o} . We assert that the r elements

$$t'_1 = t_1, \quad t'_2 = F_2\left(\frac{t_2}{t_1}\right), \quad \dots, \quad t'_r = F_r\left(\frac{t_r}{t_1}\right)$$

are uniformizing parameters at P^* , i.e. t'_1, t'_2, \dots, t'_r generate in $Q(P^*)$ the ideal of non-units. It is sufficient to show that the ideal $\mathfrak{o}^*(t'_1, t'_2, \dots, t'_r)$ is the intersection of prime zero-dimensional ideals (one of these ideals will have to be the ideal defined by the point P^* , since $t'_i = 0$ at P^*). Now this ideal is contained in the prime $(r-1)$ -dimensional ideal $\mathfrak{o}^*t_1 (= \mathfrak{o}^*\mathfrak{p}$; the center of the \mathfrak{p} -adic divisor on V^*). Modulo this prime ideal the elements $t_2/t_1, \dots, t_r/t_1$ are algebraically independent, while the residues of ξ_1, \dots, ξ_n are c_1, \dots, c_n respectively. Therefore, if we pass to the ring $\mathfrak{o}^*/\mathfrak{o}^*t_1$, we see immediately that the above assertion is equivalent to the following statement: if z_2, \dots, z_r are indeterminates over Δ , then the polynomials $f_2(z_2), \dots, f_r(z_r)$ generate in the polynomial ring $\Delta[z_2, \dots, z_r]$ an ideal which is the intersection of prime zero-dimensional ideals. Since the polynomials $f_i(z_i)$ are irreducible over Δ and since Δ is of characteristic zero, this statement is true and its proof is straightforward.

II. RESOLUTION OF THE SINGULARITIES OF AN ALGEBRAIC SURFACE

5. Two lemmas

If W is an arbitrary collection of points on a given irreducible algebraic variety V (for instance, if W is an algebraic subvariety of V), we denote by

$N(W)$ the set of all zero-dimensional valuations B (of the field Σ of rational functions on V) such that the center of B on V is in W .

LEMMA 1. *If W is an algebraic subvariety of V and if for any point P of W there exists a resolving system for $N(P)$, then there also exists a resolving system of $N(W)$.*

PROOF. Let $s = \text{dimension } W$, i.e. s is the highest dimension of the irreducible components of W . For $s = 0$ the lemma is trivial, because W consists then of a finite number of points. We therefore assume that the lemma is true for $s = \rho - 1$ and we prove that it is also true if $\dim W = \rho$.

We fix a point P_i on each irreducible ρ -dimensional component of W and we consider a resolving system V_1, \dots, V_h of $N(P_1, P_2, \dots)$. Let V^* be the join of V, V_1, \dots, V_h . The points of V^* to which there correspond singular points of V_i form an algebraic subvariety W_i^* of V^* ($i = 1, 2, \dots, h$). Let W^* be the intersection of W_1^*, \dots, W_h^* . The points of W which correspond to points of W^* form an algebraic subvariety W_1 of W , of dimension $< \rho$, since $P_i \notin W_1$. It is clear that (V_1, \dots, V_h) is also a resolving system of $N(W - W_1)$. Hence, if there exists a finite resolving system for $N(W_1)$, this system, together with V_1, \dots, V_h , will give a resolving system for $N(W)$. Since $\dim W_1 < \rho$, our induction is complete.

The second lemma deals with algebraic surfaces. Let F be a normal surface, and let P be a point of F . We apply to F a quadratic transformation of center P , getting a surface F'_1 . If F'_1 is not normal, we pass from F'_1 to a corresponding derived normal surface F_1 (see "Results," p. 290). The birational transformation which consists in passing from a given variety to a corresponding derived normal variety shall be referred to in the sequel as a *normal transformation*. We know from section 3 that the transform of P on F'_1 is a pure $(r - 1)$ -dimensional subvariety of F'_1 . We take an arbitrary point P'_1 of this subvariety. To P'_1 there will correspond on F_1 at most a finite number of points. Let P_1 be one of these points. We now repeat the above procedure, starting with the normal surface F_1 and with the point P_1 . We will get first a quadratic transform F'_2 of F_1 (with P_1 as center of the quadratic transformation) and then a derived normal variety F_2 of F'_2 . On F'_2 we select at random a point P'_2 which corresponds to P_1 , and then we let P_2 be one of the points of F_2 which corresponds to P'_2 . In this fashion we proceed indefinitely, getting an infinite sequence of models $F; F'_1, F_1; F'_2, F_2; \dots; F'_i, F_i; \dots$, and of points $P; P'_1, P_1; P'_2, P_2; \dots; P'_i, P_i; \dots$, where: a) $P'_i \in F'_i, P_i \in F_i$; b) the quotient ring $Q(P_i)$ is integrally closed and c) $Q(P) \subset Q(P'_1) \subseteq Q(P_1) \subset Q(P'_2) \subseteq Q(P_2) \dots$.

LEMMA 2. *The union of the quotient rings $Q(P_i)$ (or $\lim Q(P_i)$) is the valuation ring of a zero-dimensional valuation of Σ .*

The proof is exactly the same as the one we gave in the case of algebraically closed ground fields ("Reduction," p. 681, Theorem 10).¹⁴

¹⁴ In that proof we selected an element ζ in $Q(P)$ which is a non-unit of $Q(P)$ but is not in the "tangential" ideal. The consideration of the tangential ideal can be omitted. We know that if \mathfrak{P} is the ideal of non-units in $Q(P)$, $\mathfrak{P} = (\eta_0, \eta_1, \dots, \eta_m)$, then the ex-

6. The existence of resolving systems

Let F be a normal surface. We shall prove that the hypothesis that the field Σ of rational functions on F does not possess a resolving system leads to a contradiction. Under this hypothesis it follows, in view of Lemma 1, that there must exist a point P on F such that $N(P)$ does not possess a resolving system. By a quadratic transformation of center P , we transform F into a surface F'_1 and into the derived normal surface F_1 of F'_1 . Let W_1 be the subvariety of F_1 whose points correspond to P . It is clear that $N(P) = N(W_1)$, and hence there must exist a point P_1 on W_1 such that $N(P_1)$ does not possess a resolving system. By repeated application of this argument, we get an infinite sequence of points $P, P_1, \dots, P_i, \dots$ of the type considered in the preceding section, such that $N(P_i)$ does not possess a resolving system, $i = 1, 2, \dots$. Let B be the zero-dimensional valuation whose valuation ring B is the union of the rings $Q(P_i)$. By the local uniformization theorem, let F^* be a model on which the center P^* of B is a simple point. We have $Q(P^*) \subset B$, hence $Q(P^*) \subset Q(P_i)$, for i sufficiently high, say $i \geq m$. Every valuation of center P_i , $i \geq m$, will have P^* as center on F^* , i.e. $N(P_i) \subset N(P^*)$. Therefore F^* is a resolving surface for $N(P_i)$, if $i \geq m$, a contradiction.

7. Proof of the fundamental theorem

Let F, F' be the pair of surfaces which constitute a resolving system for a given set N of zero-dimensional valuations of our field Σ . The birational correspondence between F and F' may have fundamental points on either surface. Our first step consists in the elimination of the fundamental points of one of the two surfaces, say of F , by applying to F a sequence of successive quadratic and normal transformations, as described in the preceding section. As center of the quadratic transformation we take a fundamental point of F , one at a time, until we have exhausted all the fundamental points of F . As a result we get some new normal surface, say F_1 . If the birational correspondence between F_1 and F' still has fundamental points on F_1 , these points must be among the points which correspond to the fundamental points of F . In that case we proceed with F_1 in the same fashion. We assert that *after a finite number of steps we will get a surface $F_i = \bar{F}$ such that the birational correspondence between \bar{F} and F' has no fundamental points on \bar{F}* . For, otherwise there would exist an infinite sequence of points $P, P_1, P_2, \dots, P_i \in F_i$, such that each point P_i is fundamental and such that $Q(P) \subset Q(P_1) \subset \dots$. Let $B = \lim Q(P_i)$, and let B be the corresponding zero-dimensional valuation (Lemma 2). Let P' be the center of B on F' . Then if i is sufficiently high we will have $Q(P_i) \supseteq Q(P')$, and this is in contradiction with our hypothesis that P_i is a fundamental point of the birational correspondence between F_i and F' (see section 3).

Since under quadratic transformations simple points go into simple points

tended ideal of \mathfrak{P} in $Q(P_1)$ is a principal ideal generated by one of the elements η_i , say by η_0 . Take as ζ the element η_0 .

(section 4) and since under normal transformations simple points are not affected at all, it follows that the pair of surfaces (\bar{F}, F') is also a resolving system of N .

Our next step is similar to the step just executed, namely we now eliminate by quadratic and normal transformations the fundamental points of F' in the birational correspondence between \bar{F}, F' . However, *this time we only eliminate those fundamental points of F' which are simple points of F'* . No singular point of F' will be affected, even if it is a fundamental point. Let \bar{F}' be the transform of F' thus obtained.

We assert that the join F^ of \bar{F} and \bar{F}' is a resolving surface for N .*

The proof is straightforward. For, let B be any valuation in the set N . Let P, P', \bar{P} and P^* be the centers of B on \bar{F}, F', \bar{F}' and F^* respectively. We have to show that P^* is a simple point.

FIRST CASE: P' is a singular point. Then \bar{P} is simple (since (\bar{F}, F') is a resolving system for N). We have $Q(\bar{P}) \supseteq Q(P')$ (since the birational correspondence between \bar{F} and F' has no fundamental points on \bar{F}) and also $Q(P') = Q(\bar{P}')$ (since the singular points of F' have not been affected in the passage from F' to \bar{F}'). Therefore $Q(\bar{P}) \supseteq Q(\bar{P}')$ and consequently¹⁵ $Q(\bar{P}) = Q(P^*)$. Since \bar{P} is a simple point, it follows that also P^* is a simple point.

SECOND CASE: P' is a simple point. If P' is not a fundamental point of the birational correspondence between \bar{F} and F' , then $Q(P') = Q(\bar{P})$ (since \bar{P} is also not a fundamental point), and moreover $Q(P') = Q(\bar{P}')$ (since the points of F' which are not fundamental in the above birational correspondence have not been affected in the passage from F' to \bar{F}'). Hence $Q(\bar{P}) = Q(\bar{P}') = Q(P^*)$, and therefore P^* is a simple point. If, on the other hand, P' is a fundamental point in the birational correspondence between \bar{F} and F' , then \bar{P}' is not fundamental for the birational correspondence between \bar{F} and \bar{F}' (because the simple fundamental points of F' are eliminated in the course of the passage from F' to \bar{F}'). Hence $Q(\bar{P}') \supseteq Q(\bar{P})$, $Q(P^*) = Q(\bar{P}')$. Since P' is simple, also \bar{P}' is simple, and consequently P^* is a simple point, q.e.d.

THE JOHNS HOPKINS UNIVERSITY

¹⁵ We make use of the following property of the join V^* of two varieties V and V' : if P^*, P and P' are corresponding points of V^*, V and V' respectively and if $Q(P) \supseteq Q(P')$, then $Q(P) = Q(P^*)$. The proof is straightforward.

A NEW HOMOLOGY THEORY. II

By W. MAYER

(Received March 27, 1942)

1. Euler relations for the new homology groups

We assume familiarity with the definitions and notations of our previous paper, "A new Homology Theory," these Annals, vol. 43, 1942, pp. 370-380.

The homomorphisms

$$(1.1) \quad C^{i+q} \rightarrow C^i, \quad i = 0, 1, 2, \dots; 0 < q < p,$$

defined by

$$(1.2) \quad F^q(K^{i+q}) = K^i$$

have the group Z_q^{i+q} as nucleus and the group B_{p-q}^i as map.

In this section the underlying simplicial system Σ is supposed to be finite and if in addition we restrict ourselves to its regular subsystem $\Sigma_{(r)}$, which restriction does not influence the homology groups, the ranks of all groups involved will be finite. Thus (1.1) yields for the ranks

$$(1.3) \quad r(C^{i+q}) = r(Z_q^{i+q}) + r(B_{p-q}^i), \quad 0 < q < p, i = 0, 1, 2, \dots$$

Hence

$$(1.4) \quad r(C^i) = r(Z_q^i) + r(B_{p-q}^{i-q})$$

is true for $i \geq q$ and will remain true for $i < q$ if the rank of a group of negative dimension is put equal to zero. This follows from

$$(1.5) \quad C^i = Z_q^i \quad \text{for } i < q \text{ (and } i = 0, \pm 1, \pm 2, \dots).$$

REMARK: The introduction of chains of negative dimensions here has as its sole purpose to dispense with the handling of sub-cases. (This device was also used in §7 of our previous paper, but differently from the way adopted here.)

Hence we have

$$(1.6) \quad r(C^i) = r(Z_q^i) + r(B_{p-q}^{i-q}), \quad 0 < q < p, \quad i = 0, \pm 1, \pm 2, \dots$$

Furthermore from

$$(1.7) \quad H_q^i = Z_q^i - B_q^i$$

defining the (q, i) -homology groups, we derive

$$(1.8) \quad r(H_q^i) = r(Z_q^i) - r(B_q^i), \quad 0 < q < p, \quad i = 0, \pm 1, \pm 2, \dots$$

Eliminating $r(Z_q^i)$ from (1.6) and (1.8) we arrive at

$$(1.9) \quad r(H_q^i) = r(C^i) - r(B_q^i) - r(B_{p-q}^{i-q}), \quad 0 < q < p, \quad i = 0, \pm 1, \pm 2, \dots$$

Here we replace q and i by $p - q$ and $i - q$ respectively, to get

$$(1.10) \quad r(H_{p-q}^{i-q}) = r(C^{i-q}) - r(B_{p-q}^{i-q}) - r(B_q^{i-p}),$$

$$0 < q < p, \quad i = 0, \pm 1, \pm 2, \dots$$

From (1.9) and (1.10) we eliminate $r(B_{p-q}^{i-q})$

$$(1.11) \quad \begin{cases} r(H_q^i) - r(H_{p-q}^{i-q}) = r(C^i) - r(C^{i-q}) - r(B_q^i) + r(B_q^{i-p}), \\ 0 < q < p, \quad i = 0, \pm 1, \pm 2, \dots \end{cases}$$

For $i = np + j$, $0 \leq j < p$ (1.11) becomes

$$(1.12) \quad r(H_q^{np+j}) - r(H_{p-q}^{np+j-q}) = r(C^{np+j}) - r(C^{np+j-q}) - r(B_q^{np+j}) + r(B_q^{(n-1)p+j}).$$

For j and q fixed ($0 \leq j < p$, $0 < q < p$) we sum the relations (1.12) for $n = 0, 1, 2, \dots$, to obtain the Euler relations

$$(1.13) \quad \sum_{n=0}^{\infty} [r(H_q^{np+j}) - r(H_{p-q}^{np+j-q})] = \sum_{n=0}^{\infty} [r(C^{np+j}) - r(C^{np+j-q})].$$

The sums appearing in (1.13) are finite since Σ is finite. (For $p = 2$, $j = q = 1$ the relation (1.13) is the Euler relation for the classical modulo 2 case.)

If we let

$$(1.14) \quad r_m = r(C^m),$$

then r_m is the number of all simplexes of dimension m , simple or not. If on the other hand α_m denotes the number of the simple simplexes of dimension m then r_m is linearly expressible in terms of $\alpha_0, \alpha_1, \dots, \alpha_m$, with integral coefficients which do not depend on the underlying simplicial system Σ . Consequently the right side of (1.13) is a linear function of $\alpha_0, \alpha_1, \dots, \alpha_N$, where N denotes the dimension of Σ (i.e. N is the largest dimension of a simple simplex of Σ), with coefficients independent of Σ .

For a Euclidean polyhedron $|\Sigma|$, which is a realization of the simplicial system Σ , the left side of (1.13) has been shown to be a topological invariant. Hence this is true for the right side, and being linear in the $\alpha_0, \alpha_1, \dots, \alpha_N$, this side can differ from the Euler number

$$(1.15) \quad E(\Sigma) = \alpha_0 - \alpha_1 + \alpha_2 - \dots + (-1)^N \alpha_N$$

only by an integral factor (§4). To determine this factor we may choose the system Σ in the simplest way, and this we do in taking the system Σ_0 composed of one zero-dimensional simplex. Here $\alpha_0 = 1$ and all the other α_i , $i > 0$, are zero. Furthermore for Σ_0

$$(1.16) \quad r_0 = r_1 = \dots = r_{p-2} = 1, \quad r_n = 0, \dots n > p-2,$$

since only regular chains are admitted. For Σ_0 the right side of (1.13) becomes

$$(1.17) \quad r_j - r_{j-q} - r_{p+j-q}$$

these being the only non-vanishing terms. Two cases arise: a) $j \neq p - 1$, and b) $j = p - 1$.

Case a, $j = 0, 1, \dots, p - 2$, splits into subcases a₁) $j \geq q$ and a₂) $j < q$:

$$a_1) \quad r_j = 1, \quad r_{j-q} = 1, \quad r_{p+j-q} = 0,$$

$$a_2) \quad r_j = 1, \quad r_{j-q} = 0, \quad r_{p+j-q} = 1 \dots j < q - 1, \\ 0 \dots j = q - 1.$$

Hence for $j \neq p - 1$ the expression (1.17) is zero except for $j = q - 1$ in which case (1.17) = $E(\Sigma_0) = 1$.

In Case b, $r_j = 0$, $r_{j-q} = r_{p-q-1} = 1$, $r_{p+j-q} = r_{2p-q-1} = 0$, since $0 < q < p$. Hence for $j = p - 1$ the expression (1.17) is (independent of q) equal to $-E(\Sigma_0) = -1$.

Thus the following formula has been proved

$$(1.18) \quad \left\{ \begin{array}{l} \sum_{n=0}^{\infty} [r(H_q^{n,p+j}) - r(H_{p-q}^{n,p+j-q})] = f_j^q E(\Sigma), \\ \text{where } f_j^q = \delta_j^{q-1} - \delta_j^{p-1}, \text{ with } \delta_i^k = \begin{array}{l} 1 \dots i = k, \\ 0 \dots i \neq k. \end{array} \end{array} \right.$$

REMARK: $f_j^q = 0$ is possible only for $j \neq q - 1$ and $j \neq p - 1$ since $q - 1 \neq p - 1$.

This formula being true for any (finite) Σ suggests first

$$(1.19) \quad r(H_q^n) = 0, \quad r(H_{p-q}^n) = 0, \quad \text{for } n \neq -1 \text{ and } n - q \neq -1, (\text{mod } p),$$

and indeed these two relations hold, as will be shown in the next section. It suffices to prove the first of them, since $n \neq -1$ and $n - q \neq -1, (\text{mod } p)$ entails $n - q \neq -1$ and $(n - q) - (p - q) \equiv n \neq -1 (\text{mod } p)$. In §3 we shall prove

$$(1.20) \quad r(H_q^{p+q-1}) \text{ and } r(H_q^{p+p-1})$$

independent of q , likewise suggested by (1.18).

In the following the simplicial system is no longer supposed to be finite.

2. Proof of the rank relations

$$(2.1) \quad r(H_q^n) = 0 \quad \text{for } n \neq -1 \text{ and } n - q \neq -1, (\text{mod } p).$$

We first prove

$$(2.2) \quad \text{For } n \neq -1, (\text{mod } p) \text{ any } (q, n)\text{-cycle is a } (p - q - 1)\text{-boundary.}$$

(Of course, for $q = p - 1$ the above statement is vacuous.)

To prove (2.2) we introduce $p - 1$ linear operators D_α , $\alpha = 0, 1, \dots, p - 2$, with

$$(2.3) \quad D_\alpha(P_1 \cdots P_{n+1}) = (-1)^\alpha \alpha! \sum_{j=1}^{n+1} (P_1 \cdots P_j^{p-\alpha-1} \cdots P_{n+1}),$$

for n -chains, $n \geq 0$ and with

$$(2.3.1) \quad D_\alpha K^n = 0 \quad \text{for } n < 0, \text{ (} K^n \text{ a chain of negative dimension).}$$

For $n \geq 0$ and $\alpha = 0, 1, \dots, p-3$, we derive from (2.3)

$$(2.4) \quad \begin{aligned} FD_\alpha(P_1 \cdots P_{n+1}) &= (-1)^{\alpha+1}(\alpha+1)! \sum_{j=1}^{n+1} (P_1 \cdots P_j^{p-\alpha-2} \cdots P_{n+1}) \\ &\quad + (-1)^\alpha \alpha! \sum_{j=1}^{n+1} \sum_{k \neq j} (P_1 \cdots P_j^{p-\alpha-1} \cdots P_{n+1})_k. \end{aligned}$$

By (2.3), (2.3.1) the last term on the right side of (2.4) equals

$$(-1)^\alpha \alpha! \sum_{k=1}^{n+1} \sum_{j \neq k} (P_1 \cdots P_j^{p-\alpha-1} \cdots P_{n+1})_k = D_\alpha F(P_1 \cdots P_{n+1}).$$

and therefore we have for $n \geq 0$ and $\alpha = 0, 1, \dots, p-3$

$$(2.5) \quad (FD_\alpha - D_\alpha F)(P_1 \cdots P_{n+1}) = D_{\alpha+1}(P_1 \cdots P_{n+1}),$$

or, as an operator relation,

$$(2.6) \quad (FD_\alpha - D_\alpha F)K^n = D_{\alpha+1}K^n$$

which holds for any n (because of (2.3.1.)) and for $\alpha = 0, 1, \dots, p-3$.

From (2.6) we derive

$$(2.7) \quad D_\alpha = \sum_{r=0}^{\alpha} (-1)^r \binom{\alpha}{r} F^{\alpha-r} D_0 F^r, \quad \alpha = 0, 1, \dots, p-2.$$

Since

$$(2.8) \quad D_{p-2}(P_1 \cdots P_{n+1}) = -(p-2)!(n+1)(P_1 \cdots P_{n+1}),$$

we get from (2.7) for $\alpha = p-2$

$$(2.9) \quad -(p-2)!(n+1)K^n = \left\{ \sum_{r=0}^{p-2} (-1)^r \binom{p-2}{r} F^{p-r-2} D_0 F^r \right\} K^n.$$

But p being prime,

$$(2.10) \quad (p-2)! \equiv 1, \quad \binom{p-2}{r} \equiv (-1)^r(r+1), \quad (\text{mod } p).$$

Hence for any n -dimensional chain K^n , $n = 0, \pm 1, \pm 2, \dots$ the relation

$$(2.11) \quad \sum_{r=0}^{p-2} (r+1) F^{p-r-2} D_0 F^r K^n = -(n+1)K^n$$

holds. For a q -cycle K^n we have $D_0 F^q K^n = D_0 F^{q+1} K^n = \dots = 0$, since then either the $F^q K^n$, $F^{q+1} K^n$, \dots vanish or they are chains of negative dimensions. The relation (2.11) then becomes

$$(2.12) \quad -(n+1)K^n = F^{p-q-1} \left\{ \sum_{r=0}^{q-1} (r+1) F^{q-r-1} D_0 F^r K^n \right\}$$

thus proving (2.2).

The term corresponding to $r = q - 1$ inside the bracket in (2.12) is $q D_0 F^{q-1} K^n$. We prove (2.1) by showing that for $n - q \not\equiv -1, (\text{mod } p)$ this term is a boundary:

$$(2.13) \quad D_0 F^{q-1} K^n = F(\quad), \text{ for } n - q \not\equiv -1, (\text{mod } p).$$

For $q - 1 > n$ (2.13) is the zero-chain by (2.3.1), and hence a boundary. Here $n \not\equiv q - 1$, thus proving (2.13) for $0 > n - q + 1$. The case $n - q + 1 = 0$ can be neglected since by assumption $n - q \not\equiv -1 (\text{mod } p)$. Thus we may assume $n - q + 1 > 0$ or

$$(2.14) \quad n \geq q.$$

Let $(P_1 \cdots P_{n+1})$ be a simplex of the q -cycle K^n ; then

$$(2.15) \quad \begin{cases} D_0 F^{q-1} (P_1 \cdots P_{n+1}) = (q-1)! D_0 \sum_{(\alpha_1 \cdots \alpha_{q-1})} (P_1 \cdots P_{n+1})_{\alpha_1 \cdots \alpha_{q-1}} \\ = (q-1)! \sum_{(\beta_1 \cdots \beta_{n-q+2})} \sum_{j=1}^{n-q+2} (P_{\beta_1} \cdots P_{\beta_j}^{p-1} \cdots P_{\beta_{n-q+2}}), \end{cases}$$

the summation running over all the $(n - q + 2)$ faces of $(P_1 \cdots P_{n+1})$. Again we have to introduce linear chain-operators Δ_p , $p = 1, \dots, p - 2$ which are defined for chains of dimensions ≥ 1 by

$$(2.16) \quad \Delta_p (P_1 \cdots P_\nu) = \sum_{[j,l]} (P_1 \cdots P_j^{p-p} \cdots P_l^{p+1} \cdots P_\nu), \quad \nu \geq 2,$$

and which transform into zero chains of lower dimension. In (2.16) j, l run over $1, 2, \dots, \nu$, (without repetition, $j \neq l$). Obviously

$$(2.17) \quad \Delta_p = \Delta_\sigma \quad \text{for } p + \sigma = p - 1$$

so that, of the operators Δ_p only $\Delta_1, \Delta_2, \dots, \Delta_{(p-1)/2}$ are distinct. Since $n - q + 2 \geq 2$ we have, similarly to (2.15),

$$(2.18) \quad \begin{aligned} \Delta_1 F^{q-1} (P_1 \cdots P_{n+1}) \\ = (q-1)! \sum_{(\beta_1 \cdots \beta_{n-q+2})} \sum_{[j,l]} (P_{\beta_1} \cdots P_{\beta_j}^{p-1} \cdots P_{\beta_l}^2 \cdots P_{\beta_{n-q+2}}). \end{aligned}$$

Here, putting $n - q + 2 = \nu$, we take the boundary

$$(2.19) \quad \begin{cases} F \Delta_1 F^{q-1} (P_1 \cdots P_{n+1}) \\ = -(q-1)! \sum_{(\beta_1 \cdots \beta_\nu)} \sum_{[j,l]} (P_{\beta_1} \cdots P_{\beta_j}^{p-2} \cdots P_{\beta_l}^2 \cdots P_{\beta_\nu}) \\ + 2(\nu-1)(q-1)! \sum_{(\beta_1 \cdots \beta_\nu)} \sum_{j=1}^{\nu} (P_{\beta_1} \cdots P_{\beta_j}^{p-1} \cdots P_{\beta_\nu}) \\ + (q-1)! \sum_{(\beta_1 \cdots \beta_\nu)} \sum_{[j,l]} \sum_{k \neq j, l} (P_{\beta_1} \cdots P_{\beta_j}^{p-1} \cdots P_{\beta_l}^2 \cdots P_{\beta_\nu})_{k..} \end{cases}$$

For $\nu = n - q + 2 \geq 3$ the last term (2.19) can be written

$$(2.20) \quad q \cdot (q-1)! \sum_{(\gamma_1 \dots \gamma_{r-1})} \sum_{[j!]} (P_{\gamma_1} \dots P_{\gamma_j}^{p-1} \dots P_{\gamma_1}^2 \dots P_{\gamma_{r-1}}),$$

the first sum running over all combinations $(\gamma_1 \dots \gamma_{r-1})$ of $1, 2, \dots, n+1$. (Such a combination $(\gamma_1 \dots \gamma_{r-1})$ is derived from a combination $(\beta_1 \dots \beta_r)$ in which all of the $\gamma_1, \dots, \gamma_{r-1}$ occur. There are $(n+1) - (\nu-1) = (q + \nu - 1) - (\nu - 1) = q$ such combinations $(\beta_1 \dots \beta_r)$.) But (2.20) = $\Delta_1 F^q(P_1 \dots P_{n+1})$. For $\nu = n - q + 2 = 2$ the last term (2.19) vanishes and so does $\Delta_1 F^q(P_1 \dots P_{n+1})$, hence (2.19) can be written

$$(2.21) \quad \left\{ \begin{aligned} (F\Delta_1 - \Delta_1 F)F^{q-1}(P_1 \dots P_{n+1}) &= 2(\nu-1)D_0 F^{q-1}(P_1 \dots P_{n+1}) \\ &- (q-1)! \sum_{(\beta_1 \dots \beta_r)} \sum_{[j!]} (P_{\beta_1} \dots P_{\beta_j}^{p-2} \dots P_{\beta_1}^2 \dots P_{\beta_r}) \end{aligned} \right.$$

For $\sigma = 2, 3, \dots, (p+1)/2$ and chains of dimensions ≥ 1 , we introduce the linear operators

$$(2.22) \quad \Lambda_\sigma(P_1 \dots P_\nu) = \sum_{[j!]} (P_1 \dots P_j^{p-\sigma} \dots P_1^\sigma \dots P_\nu), \quad \nu \geq 2.$$

(Obviously $\Lambda_{(p-1)/2} = \Lambda_{(q+1)/2}$.) Now (2.21) can be written

$$(2.23) \quad (F\Delta_1 - \Delta_1 F)F^{q-1}K^n = 2(\nu-1)D_0 F^{q-1}K^n - \Lambda_2 F^{q-1}K^n.$$

Since $\nu - 1 = n - q + 1 \not\equiv 0 \pmod{p}$ by assumption, (2.13) is true if and only if

$$(2.24) \quad \Lambda_2 F^{q-1}K^n = F(\quad).$$

This follows from (2.23), K^n being a q -cycle.

For $\rho = 2, 3, \dots, (p-1)/2$ we derive equations analogous to (2.19),

$$(2.25) \quad \left\{ \begin{aligned} &F\Delta_\rho F^{q-1}(P_1 \dots P_{n+1}) \\ &= -\rho(q-1)! \sum_{(\beta_1 \dots \beta_r)} \sum_{[j!]} (P_{\beta_1} \dots P_{\beta_j}^{p-\rho-1} \dots P_{\beta_1}^{\rho+1} \dots P_{\beta_r}) \\ &\quad + (\rho+1)(q-1)! \sum_{(\beta_1 \dots \beta_r)} \sum_{[j!]} (P_{\beta_1} \dots P_{\beta_j}^{p-\rho} \dots P_{\beta_1}^\rho \dots P_{\beta_r}) \\ &\quad + (q-1)! \sum_{(\beta_1 \dots \beta_r)} \sum_{[j!]} \sum_{k \neq j, l} (P_{\beta_1} \dots P_{\beta_j}^{p-\rho} \dots P_{\beta_1}^{\rho+1} \dots P_{\beta_r})_k, \end{aligned} \right.$$

where the last term on the right is $\Delta_\rho F^q(P_1 \dots P_{n+1})$. Hence for the q -cycle K^n we have for $\rho = 2, \dots, (p-1)/2$.

$$(2.26) \quad F\Delta_\rho F^{q-1}K^n = -\rho\Lambda_{\rho+1}F^{q-1}K^n + (\rho+1)\Lambda_\rho F^{q-1}K^n.$$

But ρ and $\rho+1$ are $\not\equiv 0 \pmod{p}$, hence from (2.26) it follows that the two chains

$$(2.27) \quad \Lambda_\rho F^{q-1}K^n, \quad \Lambda_{\rho+1} F^{q-1}K^n, \quad \rho = 2, 3, \dots, (p-1)/2$$

are simultaneously bounding or not bounding.

For $\rho = (p - 1)/2$ the chains (2.27) coincide and (2.26) becomes

$$(2.28) \quad F\Delta_{(p-1)/2}F^{q-1}K^n = \Lambda_{(p-1)/2}F^{q-1}K^n.$$

Hence all the $\Lambda_p F^{q-1}K^n$ bound, $\rho = (p - 1)/2, \dots, 3, 2$, and then (2.13) follows from (2.23), which proves (2.1).

3. Proof of the rank relations

$$(3.1) \quad r(H_{p-1}^{p+p-2}) = r(H_q^{p+q-1})$$

and

$$(3.2) \quad r(H_1^{p+p-1}) = r(H_q^{p+p-1}), \quad q = 1, \dots, p-1, \nu = 0, 1, 2, \dots.$$

We first define a homomorphism

$$(3.3) \quad H_r^n \rightarrow H_s^n, \quad r < s,$$

of the (r, n) -homology group into the (s, n) -homology-group of Σ . Let

$$(3.4) \quad \{Z^n\}_{B_r^n}$$

denote the homology class of the r -cycle Z^n in H_r^n . Then either $r > n$ or $F'(Z^n) = 0^{n-r}$, hence either $s > n$ or $F^s(Z^n) = F^{s-r}F'(Z^n) = 0^{n-s}$ and Z^n is an s -cycle. The mapping of homology classes

$$(3.5) \quad \{Z^n\}_{B_r^n} \rightarrow \{Z^n\}_{B_s^n}$$

then defines the homomorphism (3.3).

REMARK: Since a $(p - r)$ -boundary of dimension n is a $(p - s)$ -boundary of this dimension, $B_r^n \subset B_s^n$, and the mapping (3.5) is independent of the choice of Z^n in the class left in (3.5).

Let us consider the nucleus of the homomorphism (3.3) as defined by (3.5): The class on the left in (3.5) belongs to the nucleus if and only if $Z^n \subset B_s^n$, i.e. if and only if for some chain K^{n+p-s}

$$(3.6) \quad F^{p-s}(K^{n+p-s}) = Z^n.$$

For $n \geq r$ from $F'(Z^n) = 0^{n-r}$ then

$$(3.7) \quad F^{p+r-s}(K^{n+p-s}) = 0^{n-r}$$

follows. For $n < r$ we have $n + p - s < r + p - s$ and again K^{n+p-s} is an $(r + p - s)$ -cycle. Thus we have proved that $\{Z^n\}_{B_s^n}$ belongs to the nucleus of the homomorphism (3.3) if and only if a $(p + r - s)$ -cycle of dimension $(p + n - s)$ exists (i.e. an element of Z_{p+r-s}^{p+n-s}) such that (3.6) holds.

But (3.6) defines a homomorphism

$$(3.8) \quad H_{p+r-s}^{p+n-s} \rightarrow H_r^n.$$

Indeed, Z^n is an r -cycle if K^{n+p-s} is a $(p + r - s)$ -cycle. Furthermore, to a $p - (p + r - s) = (s - r)$ -boundary K^{n+p-s} ,

$$(3.9) \quad K^{n+p-s} = F^{s-r}(K^{n+p-r})$$

corresponds in (3.6)

$$(3.10) \quad Z^n = F^{p-s}F^{s-r}(K^{n+p-r}) = F^{p-r}(K^{n+p-r}),$$

i.e. a $(p-r)$ -boundary. Our above result can be stated as follows:

$\{Z^n\}_{B_r^n}$ belongs to the nucleus of the homomorphism (3.5) if and only if it belongs to the map-group of the homomorphism (3.8). Or: *The nucleus of (3.5) is the map of (3.8).*

We now consider the nucleus of the homomorphism (3.8) as defined by (3.6) or by

$$(3.11) \quad \{K^{p+n-s}\}_{B_{p+r-s}^{p+n-s}} \rightarrow \{F^{p-s}(K^{p+n-s})\}_{B_r^n}.$$

The class on the left belongs to the nucleus of (3.8) if and only if

$$(3.12) \quad F^{p-s}(K^{n+p-s}) = F^{p-r}(K_1^{n+p-r})$$

for some chain K_1^{n+p-r} . But then

$$(3.13) \quad F^{p-s}[K^{n+p-s} - F^{s-r}(K_1^{n+p-r})] = 0^n,$$

and the $(p-s)$ -cycle $K^{n+p-s} - F^{s-r}(K_1^{n+p-r})$, since $p - (p+r-s) = s-r$, belongs to the class on the left in (3.11).

Thus the class on the left in (3.11) belongs to the nucleus of the homomorphism (3.8) if and only if this class contains a $(p-s)$ -cycle.

If K^{n+p-s} is a $(p-s)$ -cycle, then in a way analogous to (3.5) (since $r+p-s > p-s$)

$$(3.14) \quad \{K^{n+p-s}\}_{B_{p+r-s}^{n+p-s}} \rightarrow \{K^{n+p-s}\}_{B_{r+p-s}^{n+p-s}}$$

defines a homomorphism

$$(3.15) \quad H_{p-s}^{n+p-s} \rightarrow H_{r+p-s}^{n+p-s}.$$

Our last result can now be stated as follows:

The class on the left in (3.11) belongs to the nucleus of (3.8) if and only if it belongs to the map of (3.15).

Or: *The nucleus of (3.8) is the map of (3.15).*

Thus, in the three homomorphisms

$$(3.16) \quad \begin{cases} H_r^n \rightarrow H_s^n \\ H_{p+r-s}^{p+n-s} \rightarrow H_r^n \\ H_{p-s}^{p+n-s} \rightarrow H_{r+p-s}^{p+n-s} \end{cases}$$

as defined by (3.5), (3.11) and (3.14) respectively, the nucleus of the first (second) is the map of the second (third) homomorphism.

Since $p + r - s > p - s$, the relations (3.16) may be continued to give the chain of homomorphisms:

$$(3.17) \quad \left\{ \begin{array}{l} H_r^n \rightarrow H_s^n \\ H_{p+r-s}^{n+p-s} \rightarrow H_r^n \\ H_{p-s}^{n+p-s} \rightarrow H_{p+r-s}^{n+p-s} \\ H_{p-r}^{n+p-r} \rightarrow H_{p-s}^{n+p-s} \\ H_{s-r}^{n+p-r} \rightarrow H_{p-r}^{n+p-r} \\ H_s^{n+p} \rightarrow H_{s-r}^{n+p-r} \\ H_r^{n+p} \rightarrow H_s^{n+p} \end{array} \right.$$

in which the nucleus of any homomorphism coincides with the map of the following one. We write for

$$(3.18) \quad s = r + 1, \quad n - r = \nu p - 1, \quad \nu = 0, 1, 2, \dots, \quad r = 1, \dots, p - 2,$$

the first three homomorphisms (3.17)

$$(3.19) \quad \begin{array}{l} H_r^{\nu p+r-1} \rightarrow H_{r+1}^{\nu p+r-1} \\ H_{p-1}^{\nu p+p-2} \rightarrow H_r^{\nu p+r-1} \\ H_{p-r-1}^{\nu p+p-2} \rightarrow H_{p-1}^{\nu p+p-2} \end{array}$$

By our earlier result (2.1) $H_{r+1}^{\nu p+r-1}$ and $H_{p-r-1}^{\nu p+p-2}$ are zero-groups. (Indeed $\nu p + r - 1 \not\equiv -1$ and $\not\equiv r \pmod{p}$ and $\nu p + p - 2 \not\equiv -1$ and $\not\equiv p - r - 2 \pmod{p}$.) Thus the nucleus of the first homomorphism and hence the map of the second is the group $H_r^{\nu p+r-1}$ itself. Furthermore, the map of the third homomorphism and hence the nucleus of the second one is the zero-subgroup of $H_{p-1}^{\nu p+p-2}$. Therefore the second homomorphism (3.19) is an isomorphism

$$(3.20) \quad H_{p-1}^{\nu p+p-2} \approx H_r^{\nu p+r-1},$$

thus proving (3.1) (first for $r = 1, 2, \dots, p - 2$. But for $r = p - 1$ (3.1) becomes an identity).

Now, beginning with the fourth, we write a second triple of homomorphisms (3.17) again using (3.18):

$$(3.21) \quad \begin{array}{l} H_{p-r}^{\nu p+p-1} \rightarrow H_{p-r-1}^{\nu p+p-2} \\ H_1^{\nu p+p-1} \rightarrow H_{p-r}^{\nu p+p-1} \\ H_{r+1}^{(\nu+1)p+r-1} \rightarrow H_1^{\nu p+p-1}. \end{array}$$

Here $H_{p-r-1}^{\nu p+p-2}$ and $H_{r+1}^{(\nu+1)p+r-1}$ are zero-groups, and the same reasoning as used before proves the isomorphism

$$(3.22) \quad H_1^{\nu p+p-1} \approx H_{p-r}^{\nu p+p-1},$$

thus proving (3.2). The ranks (3.1) are in general different from the ranks (3.2), as follows from the Euler relations (1.18) since, in general, $E(\Sigma) \neq 0$.

4. The Euler number

THEOREM (4.1): *If α_i denotes the number of i -dimensional simple simplexes of the finite simplicial system Σ of dimension n then (up to a factor) the Euler number*

$$E(\Sigma) = \sum_{i=0}^n (-1)^i \alpha_i$$

is the only invariant of Σ with respect to its subdivisions, which is linear in $\alpha_0, \alpha_1, \dots, \alpha_n$.

REMARK: In case Σ is a polyhedron, $E(\Sigma)$ (up to a factor) will be the only topological invariant linear in $\alpha_1, \dots, \alpha_n$.

Let $\sum A_k \alpha_k$, $k = 0, 1, \dots, n$, be an invariant with respect to subdivisions of simplicial systems Σ of dimension n . Then we prove

$$(4.2) \quad \sum_{k=0}^n A_k \alpha_k = fE(\Sigma).$$

To prove (4.2) we can choose any system Σ of dimension n since by assumption the A_k do not depend on Σ . We choose for Σ the system composed of the n -simplex $(P_1 \cdots P_{n+1})$ and its faces denoted by Σ_n . Here

$$(4.3) \quad \alpha_k(\Sigma_n) = \alpha_k^n = \binom{n+1}{k+1}.$$

Let Σ_n be subdivided into Σ'_n and let $\alpha_k'^n$ be the number of the k -simplexes of Σ'_n . Then

$$(4.4) \quad \sum_{k=0}^n A_k (\alpha_k'^n - \alpha_k^n) = 0,$$

i.e. the "vector" $\xi_k^n = \alpha_k'^n - \alpha_k^n$, $k = 0, 1, \dots, n$, lies in the hyperplane $\sum A_k \xi_k = 0$. If for suitable choices of such subdivisions we arrive at n such linearly-independent vectors ξ_k^n , $\nu = 1, 2, \dots, n$, then the A_k 's are determined up to a factor f and (4.2) is proved.

We choose the subdivision Σ'_n of Σ_n yielding ξ_k^n , $\nu = 1, \dots, n$ (the ν th vector in our construction) by introducing the centroid of a ν -face of Σ_n .

Let $(P_1 \cdots P_{\nu+1})$ be that ν -face, $\nu = 0$ included, S its centroid and α_k^n the number of k -simplexes of Σ'_n . A k -simplex of Σ'_n not containing S is of the type $(P_{\alpha_1} \cdots P_{\alpha_{k+1}})$ but not of the type $(P_1 \cdots P_{\nu+1} P_{\beta_1} \cdots P_{\beta_{k-\nu}})$, which (for $\nu \neq 0$) does not exist in Σ'_n .

Hence the number of k -simplexes of Σ'_n not containing S is

$$(4.5.1) \quad \binom{n+1}{k+1} - \binom{n-\nu}{k-\nu}.$$

A k -simplex of Σ_n^* containing S is of the type $(SP_{\alpha_1} \cdots P_{\alpha_k})$ but not of the type $(SP_1 \cdots P_{r+1}P_{\beta_1} \cdots P_{\beta_{k-r-1}})$.

The number of k -simplexes of Σ_n^* with one vertex S therefore is

$$(4.5.2) \quad \binom{n+1}{k} - \binom{n-\nu}{k-\nu-1}.$$

Thus

$$(4.6) \quad \alpha_k^{n,\nu} = \binom{n+1}{k+1} + \binom{n+1}{k} - \binom{n-\nu}{k-\nu} - \binom{n-\nu}{k-\nu-1}$$

and

$$(4.7) \quad \xi_k^{n,\nu} = \alpha_k^{n,\nu} - \alpha_k^n = \binom{n+1}{k} - \binom{n-\nu}{k-\nu} - \binom{n-\nu}{k-\nu-1},$$

$$\nu = 0, 1, \dots, n+1.$$

For $\xi_k^{n,\nu}$ the relations

$$(4.8) \quad \sum_{k=0}^n (-1)^k \xi_k^{n,\nu} = 0$$

hold. This of course is a consequence of the invariance of the Euler number, but is readily verified as follows. The left side of (4.8)

$$\begin{aligned} &= \sum_{k=0}^n (-1)^k \binom{n+1}{k} + \sum_{k=1}^{n+1} (-1)^k \binom{n-\nu}{k-\nu-1} - \sum_{k=0}^n (-1)^k \binom{n-\nu}{k-\nu-1} \\ &= (1-1)^{n+1} - (-1)^{n+1} + (-1)^{n+1} = 0. \end{aligned}$$

Furthermore,

$$(4.9) \quad \xi_k^{n,0} = \binom{n+1}{k} - \binom{n}{k} - \binom{n}{k-1} = 0$$

and

$$(4.10) \quad \xi_0^{n,\nu} = \binom{n+1}{0} - \binom{n-\nu}{-\nu} - \binom{n-\nu}{-\nu-1} = \begin{matrix} 0 & \cdots & \nu=0, \\ 1 & \cdots & \nu \neq 0. \end{matrix}$$

From (4.7) for $n, \nu > 0$ we get

$$\begin{aligned} (4.11) \quad \xi_k^{n-1,\nu-1} &= \binom{n}{k} - \binom{n-\nu}{k-\nu+1} - \binom{n-\nu}{k-\nu} \\ &= \xi_{k+1}^{n,\nu} - \binom{n+1}{k+1} + \binom{n}{k} = \xi_{k+1}^{n,\nu} - \binom{n}{k+1}, \end{aligned}$$

or

$$(4.12) \quad \xi_{k+1}^{n,\nu} = \xi_k^{n-1,\nu-1} + \binom{n}{k+1}.$$

We are now able to prove the linear independence of the vectors $\xi_k^{n,\nu}$, $\nu = 1, \dots, n$, ($k = 0, 1, \dots, n$). From

$$(4.13) \quad \sum_{\nu=1}^n a_{\nu} \xi_k^{n,\nu} = 0$$

by (4.10) for $k = 0$ we deduce

$$(4.14) \quad \sum_{\nu=1}^n a_{\nu} = 0,$$

and then by (4.12) and (4.9) for $k = 1, \dots, n$

$$(4.15) \quad \sum_{\nu=1}^n a_{\nu} \xi_{k-1}^{n-1,\nu-1} = \sum_{\nu=0}^{n-1} a_{\nu+1} \xi_{k-1}^{n-1,\nu} = \sum_{\nu=1}^{n-1} a_{\nu+1} \xi_{k-1}^{n-1,\nu} = 0.$$

Thus from the linear independence of the $(n-1)$ vectors $\xi_k^{n-1,\nu}$, $\nu = 1, \dots, n-1$ ($k = 0, \dots, n-1$) (by (4.13), (4.14) and (4.15)) the linear independence of the n vectors $\xi_k^{n,\nu}$, $\nu = 1, \dots, n$, ($k = 0, 1, \dots, n$), follows. But $\xi_k^{1,1} = 1$, $k = 0, 1$, hence, of the sets of vectors $\xi_k^{\rho,\nu}$, $\rho = 2, \dots, n$, $\nu = 1, \dots, \rho$, ($k = 0, 1, \dots, \rho$) each one is a linearly independent set, thus proving (4.1).

ITERATION OF ANALYTIC FUNCTIONS

BY CARL LUDWIG SIEGEL

(Received April 23, 1942)

Let

$$(1) \quad f(z) = \sum_{k=1}^{\infty} a_k z^k$$

be a power series without constant term and denote by $R > 0$ its radius of convergence. The fixed point $z = 0$ of the mapping $z \rightarrow f(z)$ is called stable, if there exist two positive finite numbers $r_0 \leq R$ and $r \leq R$, such that for all points z of the circle $|z| < r_0$ the set of image points $z_1 = f(z)$, $z_{n+1} = f(z_n)$ ($n = 1, 2, \dots$) lies in the circle $|z| < r$.

It is easy to prove the stability in the case $|a_1| < 1$, for then a positive number $r_0 < R$ exists, such that the inequality $|f(z)| \leq |z|$ holds for $|z| < r_0$, and $r = r_0$ has the required property. Henceforth, the inequality $|a_1| \geq 1$ is assumed.

If $z = 0$ is stable, then the images z_n ($n = 1, 2, \dots$) of the points z of the circle $|z| < r_0$ under the mapping $z \rightarrow f(z)$ and its iterations cover a domain D which is connected and contains the point $z = 0$. For all z in D , the image point $f(z)$ again lies in D . Let

$$(2) \quad z = \varphi(\zeta) = \zeta + \sum_{k=2}^{\infty} c_k \zeta^k$$

be the power series mapping a certain circle $|\zeta| < \rho$ of the ζ plane conformally onto the universal covering surface of D . Then the formula

$$\varphi(\zeta) = z \rightarrow f(z) = z_1 = \varphi(\zeta_1)$$

defines a function $\zeta_1 = g(\zeta)$ which is regular in the circle $|\zeta| < \rho$ and satisfies there the inequality $|g(\zeta)| < \rho$; moreover $g(0) = 0$ and $g'(0) = 1$. It follows from Schwarz's lemma that $|a_1| = 1$ and $\zeta_1 = a_1 \zeta$. Consequently, the functional equation of Schröder

$$(3) \quad \varphi(a_1 \zeta) = f(\varphi(\zeta))$$

has a convergent solution $\varphi(\zeta) = \zeta + \dots$.

On the other hand, it is obvious that $z = 0$ is stable, if $|a_1| = 1$ and the functional equation (3) has a convergent solution.

If a_1 is an n^{th} root of unity, then $z = 0$ is stable, if and only if the $(n - 1)^{\text{th}}$ iteration of the mapping $z \rightarrow f(z)$ is the identity. This is also easily proved by direct calculation. We assume now that $|a_1| = 1$ and $a_1^n \neq 1$ for $n = 1, 2, \dots$.

By (1), (2) and (3),

$$(4) \quad \sum_{k=2}^{\infty} c_k (a_1^k - a_1) \zeta^k = \sum_{l=2}^{\infty} a_l \left(\zeta + \sum_{r=2}^{\infty} c_r \zeta^r \right)^l;$$

hence c_k ($k = 2, 3, \dots$) is a polynomial in c_2, \dots, c_{k-1} whose coefficients depend upon a_1, \dots, a_k , and there exists exactly one formal (convergent or divergent) solution $\varphi(\zeta) = \zeta + \dots$ of (3). The first example of a convergent series $f(z) = a_1 z + \dots$ with *divergent* Schröder series $\varphi(\zeta)$ has been given by Pfeiffer.¹ Later Cremer² has constructed such examples for arbitrary a_1 satisfying the condition

$$\liminf_{n \rightarrow \infty} |\dot{a}_1^n - 1|^{1/n} = 0.$$

These a_1 are very closely approximated by certain roots of unity, and their linear Lebesgue measure on the unit circle $|a_1| = 1$ is 0.

Until now, however, it was not known if there exists a number a_1 of absolute value 1, such that every convergent power series $f(z) = a_1 z + \dots$ has a *convergent* Schröder series. Julia³ tried to prove the erroneous hypothesis that the Schröder series is always divergent, if $f(z) - a_1 z$ is a rational function and not identically 0. We shall demonstrate the following

THEOREM: *Let*

$$(5) \quad \log |a_1^n - 1| = O(\log n) \quad (n \rightarrow \infty);$$

then the Schröder series is convergent.

Write $a_1 = e^{2\pi i \omega}$; then the condition (5) may be expressed in the form

$$\left| \omega - \frac{m}{n} \right| > \lambda n^{-\mu},$$

for arbitrary integers m and n , $n \geq 1$, where λ and μ denote positive numbers depending only upon ω . It is easily seen that (5) holds for all points of the unit circle $|a_1| = 1$ with the exception of a set of measure 0.

LEMMA 1: *Let x_p ($p = 1, \dots, r$) and y_q ($q = 1, \dots, s$) be positive integers, $r \geq 0, s \geq 2, r + s = t$,*

$$\sum_{p=1}^r x_p + \sum_{q=1}^s y_q = k, \quad \sum_{q=1}^s y_q > \frac{k}{2}, \quad y_q \leq \frac{k}{2} \quad (q = 1, \dots, s);$$

then

$$(6) \quad \prod_{p=1}^r x_p \prod_{q=1}^s y_q^2 \geq k^3 8^{t-t}.$$

PROOF: Denote the left-hand side of (6) by L and consider first the case $k < 2t - 2$. Then

$$(7) \quad k^{-3} L \geq k^{-3} > (2t - 2)^{-3}.$$

¹ G. A. Pfeiffer, *On the conformal mapping of curvilinear angles. The functional equation $\varphi[f(x)] = a_1 \varphi(x)$* , Trans. Amer. Math. Soc. 18, pp. 185-198 (1917).

² H. Cremer, *Über die Häufigkeit der Nichtzentren*, Math. Ann. 115, pp. 573-580 (1938).

³ G. Julia, *Sur quelques problèmes relatifs à l'itération des fractions rationnelles*, C. R. Acad. Sci. Paris 168, pp. 147-149 (1919).

Assume now $k \geq 2t - 2$ and let

$$\left[\frac{k}{2} \right] = g, \quad r + \sum_{q=1}^s y_q = \eta.$$

Then

$$t \leq g + 1 \leq g + 1 + r \leq \eta \leq k, \quad \sum_{p=1}^r x_p = k - \eta + r,$$

whence

$$\prod_{p=1}^r x_p \geq k - \eta + 1, \quad \prod_{q=1}^s y_q \geq \begin{cases} \eta - t + 1, & \text{if } \eta \leq g - 1 + t \\ (\eta - g - t + 2)g, & \text{if } \eta \geq g - 1 + t. \end{cases}$$

In the interval $g + 1 \leq \eta \leq g - 1 + t$,

$$(k - \eta + 1)(\eta - t + 1)^2 \geq \min \{ (k - g)(g - t + 2)^2, (k - g - t + 2)g^2 \};$$

in the interval $g - 1 + t \leq \eta \leq k$,

$$(k - \eta + 1)(\eta - g - t + 2)^2 g^2 \geq (k - g - t + 2)g^2;$$

in the interval $0 \leq \xi \leq g$,

$$(k - g)(g - \xi)^2 - (k - g - \xi)g^2 = \{ (k - g)\xi - (2k - 3g)g \} \xi \leq g(2g - k)\xi \leq 0;$$

consequently

$$(8) \quad L \geq (k - g)(g - t + 2)^2$$

$$k^{-3}L \geq \frac{k - g}{k} \left(\frac{g - t + 2}{k} \right)^2 \geq \frac{1}{2}(2t - 2)^{-2} \geq (2t - 2)^{-3}.$$

Now

$$t - 1 \leq 2^{t-2} \quad (t = 2, 3, \dots),$$

and the lemma follows from (7) and (8).

We use the abbreviation

$$\epsilon_n = |a_1^n - 1|^{-1} \quad (n = 1, 2, \dots).$$

On account of (5), the inequalities

$$\epsilon_n < (2n)^r \quad (n = 1, 2, \dots)$$

are fulfilled for a certain constant positive value r . We define

$$N_1 = 2^{2r+1}, \quad N_2 = 8^r N_1 = 2^{5r+1}.$$

LEMMA 2: Let m_l ($l = 0, \dots, r$) be integral, $r \geq 0$ and $m_0 > m_1 > \dots > m_r > 0$; then

$$(9) \quad \prod_{l=0}^r \epsilon_{m_l} < N_1^{r+1} \left\{ m_0 \prod_{l=1}^r (m_{l-1} - m_l) \right\}^r.$$

PROOF: The assertion is true in the case $r = 0$; assume $r > 0$ and apply induction.

We have the identity

$$a_1^q(a_1^{p-q} - 1) = (a_1^p - 1) - (a_1^q - 1) \quad (0 < q < p),$$

whence

$$\epsilon_{p-q}^{-1} \leq \epsilon_p^{-1} + \epsilon_q^{-1}$$

$$\min(\epsilon_p, \epsilon_q) \leq 2\epsilon_{p-q} < 2^{r+1}(p-q)^r.$$

This simple remark is the main argument of the whole proof.

Let ϵ_{m_l} ($l = 0, \dots, r$) have its minimum value for $l = h$. Then

$$(10) \quad \epsilon_{m_h} < 2^{r+1} \min \{(m_{h-1} - m_h)^r, (m_h - m_{h+1})^r\},$$

if we define moreover $m_{-1} = \infty$ and $m_{r+1} = -\infty$. On the other hand, the lemma being true for $r-1$ instead of r , we have

$$(11) \quad \epsilon_{m_h}^{-1} \prod_{l=0}^r \epsilon_{m_l} < N_1^r \left\{ \frac{m_0(m_{h-1} - m_{h+1})}{(m_{h-1} - m_h)(m_h - m_{h+1})} \prod_{l=1}^r (m_{l-1} - m_l) \right\}^r.$$

Since

$$\frac{m_{h-1} - m_{h+1}}{(m_{h-1} - m_h)(m_h - m_{h+1})} = \frac{1}{m_{h-1} - m_h} + \frac{1}{m_h - m_{h+1}} \leq \frac{2}{\min(m_{h-1} - m_h, m_h - m_{h+1})}$$

the inequality (9) follows from (10) and (11).

Consider now the sequence of positive numbers $\delta_1 = 1, \delta_2, \delta_3, \dots$ recurrently defined in the following way: For every $k > 1$, let μ_k denote the maximum of all products $\delta_{l_1} \delta_{l_2} \dots \delta_{l_r}$ with $l_1 + l_2 + \dots + l_r = k > l_1 \geq l_2 \geq \dots \geq l_r \geq 1$, $2 \leq r \leq k$, and put

$$(12) \quad \delta_k = \epsilon_{k-1} \mu_k.$$

LEMMA 3:

$$(13) \quad \delta_k \leq k^{-2\nu} N_2^{k-1} \quad (k = 1, 2, \dots).$$

PROOF: The assertion is true in the case $k = 1$; assume $k > 1$ and apply induction.

The numbers $\alpha_k = k^{-2\nu} N_2^{k-1}$ satisfy the inequalities

$$\frac{\alpha_k \alpha_l}{\alpha_{k+l}} = (k^{-1} + l^{-1})^{2\nu} N_2^{-1} \leq 2^{2\nu} N_2^{-1} < 1 \quad (k \geq 1, l \geq 1),$$

and consequently

$$(14) \quad \delta_{j_1} \delta_{j_2} \dots \delta_{j_f} \leq j^{-2\nu} N_2^{j-1} \quad (1 \leq j_1 + \dots + j_f = j < k; f \geq 1).$$

By (12), there exists a decomposition

$$\delta_k = \epsilon_{k-1} \delta_{g_1} \delta_{g_2} \dots \delta_{g_a} \quad (g_1 + \dots + g_a = k > g_1 \geq \dots \geq g_a \geq 1).$$

In the case $g_1 > k/2$, we use this formula with g_1 instead of k and find a decomposition

$$\delta_{v_1} = \epsilon_{v_1-1} \delta_{h_1} \delta_{h_2} \cdots \delta_{h_\beta} \quad (h_1 + \cdots + h_\beta = g_1 > h_1 \geq \cdots \geq h_\beta \geq 1);$$

if also $h_1 > k/2$, we decompose again

$$\delta_{h_1} = \epsilon_{h_1-1} \delta_{i_1} \delta_{i_2} \cdots \delta_{i_\gamma} \quad (i_1 + \cdots + i_\gamma = h_1 > i_1 \geq \cdots \geq i_\gamma \geq 1),$$

and so on. Writing $k_0 = k$, $k_1 = g_1$, $k_2 = h_1$, \cdots , we obtain in this manner the formula

$$\delta_k = \prod_{p=0}^r (\epsilon_{k_p-1} \Delta_p)$$

with $k = k_0 > k_1 > \cdots > k_r > k/2$, where Δ_p denotes for $p = 0, \cdots, r$ a certain product $\delta_{j_1} \cdots \delta_{j_f}$ and

$$j_1 + \cdots + j_f = \begin{cases} k_p - k_{p+1} & (p = 0, \cdots, r-1) \\ k_r & (p = r), \end{cases}$$

all subscripts j_1, \cdots, j_f being $\leq k/2$. The number f depends upon p ; let $f = s$ for $p = r$.

Using (13) for the s single factors of Δ_r and applying (14) for the estimation of Δ_p ($p = 0, \cdots, r-1$), we find the inequality

$$\prod_{p=0}^r \Delta_p \leq N_2^{k-r-s} \left\{ \prod_{q=1}^s j_q \prod_{p=1}^r (k_{p-1} - k_p) \right\}^{-2p},$$

where $1 \leq j_q \leq k/2$ ($q = 1, \cdots, s$) and $j_1 + \cdots + j_s = k_r$. By Lemma 2,

$$\prod_{p=0}^r \epsilon_{k_p-1} < N_1^{r+1} \left\{ k \prod_{p=1}^r (k_{p-1} - k_p) \right\}^r,$$

and consequently

$$\delta_k < N_1^{r-1} N_2^{k-t} \left(k^{-1} \prod_{p=1}^r x_p \prod_{q=1}^s y_q^2 \right)^{-r}$$

with $t = r + s$, $x_p = k_{p-1} - k_p$, $y_q = j_q$. By Lemma 1,

$$N_2^{1-k} k^{2r} \delta_k < N_1^{r+1} N_2^{1-t} 8^{r(t-1)} \leq \left(\frac{8^r N_1}{N_2} \right)^{t-1} = 1,$$

and (13) is proved.

PROOF OF THE THEOREM: Since the power series (1) has a positive radius of convergence, there exists a positive number a , such that $|a_n| \leq a^{n-1}$ ($n = 2, 3, \cdots$). The functional equation (3) remains true under the transformation $f(z) \rightarrow af(z/a)$, $\varphi(\zeta) \rightarrow a\varphi(\zeta/a)$; hence we may assume $|a_n| \leq 1$ ($n = 2, 3, \cdots$).

Instead of (4), we consider the functional equation

$$(15) \quad \sum_{k=2}^{\infty} \eta_k \gamma_k \zeta^k = \sum_{l=2}^{\infty} \left(\zeta + \sum_{r=2}^{\infty} \gamma_r \zeta^r \right)^l,$$

where η_2, η_3, \dots are positive parameters. Then the coefficients $\gamma_1 = 1, \gamma_2, \gamma_3, \dots$ are uniquely determined by the formula

$$(16) \quad \gamma_k = \eta_k^{-1} \sum \gamma_{l_1} \gamma_{l_2} \cdots \gamma_{l_r} \quad (k = 2, 3, \dots),$$

where l_1, \dots, l_r run over all positive integral solutions of $l_1 + \dots + l_r = k$ ($r = 2, \dots, k$). Write $\gamma_k = \sigma_k$ in the case $\eta_k = \epsilon_{k-1}^{-1}$ ($k = 2, 3, \dots$), and $\gamma_k = \tau_k$ in the case $\eta_k = 1$.

The inequality

$$(17) \quad \sigma_k \leq \delta_k \tau_k$$

is true for $k = 1$. Applying induction, we infer from (12) and (16) that

$$\sigma_k \leq \epsilon_{k-1} \mu_k \sum \tau_{l_1} \tau_{l_2} \cdots \tau_{l_r} = \delta_k \tau_k;$$

hence (17) holds for all values of k .

On the other hand, the power series

$$\psi = \sum_{k=1}^{\infty} \tau_k \zeta^k$$

satisfies the equation

$$\psi - \zeta = (1 - \psi)^{-1} \psi^2,$$

whence

$$4\psi = 1 + \zeta - (1 - 6\zeta + \zeta^2)^{\frac{1}{2}};$$

consequently ψ converges in the circle $|\zeta| < 3 - 2\sqrt{2}$.

By (4), (15) and (17),

$$|c_k| \leq \delta_k \tau_k \quad (k = 2, 3, \dots).$$

It follows now from Lemma 3, that the Schröder series $\varphi(\zeta)$ converges in the circle $|\zeta| < (3 - 2\sqrt{2})2^{-6r-1}$.

NOTE ON AUTOMORPHIC FUNCTIONS OF SEVERAL VARIABLES

BY CARL LUDWIG SIEGEL

(Received April 28, 1942)

1

Some years ago I found a method¹ of estimating the number of linearly independent modular forms of degree n and of weight g , which has been useful for the demonstration² of certain identities in the analytical theory of quadratic forms. The object of this note is to prove an analogous estimate concerning automorphic functions.

Let $\mathfrak{Z} = (z_{ki})$ be a complex symmetric matrix with n rows, and consider the space E defined by the condition $\mathfrak{E} - \mathfrak{Z}\bar{\mathfrak{Z}} > 0$, with the line element

$$ds = \sigma^{\frac{1}{2}} \{ d\mathfrak{Z}(\mathfrak{E} - \mathfrak{Z}\bar{\mathfrak{Z}})^{-1} d\bar{\mathfrak{Z}}(\mathfrak{E} - \mathfrak{Z}\bar{\mathfrak{Z}})^{-1} \},$$

the symbol σ denoting the trace. If \mathfrak{A} and \mathfrak{B} are n -rowed complex square matrices satisfying $\mathfrak{A}\mathfrak{B}' = \mathfrak{B}\mathfrak{A}'$ and $\mathfrak{A}\bar{\mathfrak{A}}' - \mathfrak{B}\bar{\mathfrak{B}}' = \mathfrak{E}$, then the linear transformation

$$(1) \quad \mathfrak{Z}^* = (\mathfrak{A}\mathfrak{Z} + \mathfrak{B})(\bar{\mathfrak{B}}\mathfrak{Z} + \bar{\mathfrak{A}})^{-1}$$

defines an isometric mapping of E onto itself. Those transformations constitute a group Ω .

Denoting by $\rho(\mathfrak{Z}_1, \mathfrak{Z}_0)$ the distance of two arbitrary points \mathfrak{Z}_1 and \mathfrak{Z}_0 of E , we have³

$$\rho(\mathfrak{Z}_1, 0) = \left(\sum_{k=1}^n u_k^2 \right)^{\frac{1}{2}},$$

where

$$u_k = \log \frac{1 + \lambda_k^{\frac{1}{2}}}{1 - \lambda_k^{\frac{1}{2}}} \quad (k = 1, \dots, n)$$

and $\lambda_1, \dots, \lambda_n$ are the characteristic roots of the hermitian matrix $\mathfrak{Z}_1\bar{\mathfrak{Z}}_1$. Since

$$\frac{4\lambda_k}{1 - \lambda_k} = e^{u_k} + e^{-u_k} - 2 = 2 \sum_{l=1}^{\infty} \frac{u_k^{2l}}{(2l)!}$$

¹ C. L. Siegel, *Einführung in die Theorie der Modulfunktionen n-ten Grades*, Math. Ann. 116, pp. 617-657 (1939).

² H. Maass, *Zur Theorie der automorphen Funktionen von n Veränderlichen*, Math. Ann. 117, pp. 538-578 (1940).

E. Witt, *Eine Identität zwischen Modulformen zweiten Grades*, Abh. Math. Sem. Hansischen Univ. 14, pp. 323-337 (1941).

H. Maass, *Modulformen und quadratische Formen über dem quadratischen Zahlkörper $R(\sqrt{5})$* , Math. Ann. 118, pp. 65-84 (1942).

³ C. L. Siegel, *Symplectic geometry*, submitted for publication in the Amer. J. Math.

and

$$\sum_{k=1}^n u_k^{2i} \leq \left(\sum_{k=1}^n u_k^2 \right)^i = \rho^{2i}(\mathfrak{Z}_1, 0),$$

we obtain the inequality

$$(2) \quad \sum_{k=1}^n \frac{\lambda_k}{1 - \lambda_k} \leq \sinh^2 \frac{1}{2} \rho, \quad \rho = \rho(\mathfrak{Z}_1, 0).$$

Let Δ be a subgroup of Ω , discontinuous in E , and assume that all frontier points of a fundamental domain F of Δ belong to E ; i.e. E is compact relative to Δ . The least upper bound of the distance $\rho(\mathfrak{Z}_1, \mathfrak{Z}_0)$ for two variable points \mathfrak{Z}_1 and \mathfrak{Z}_0 of F is a finite positive number δ , the diameter of F . We use the abbreviations

$$(3) \quad \nu = \frac{n(n+1)}{2}, \quad b = \sinh^2 \frac{1}{2} \delta, \quad c = (\nu + 1)b'.$$

2

An analytic function $f(\mathfrak{Z})$ of the ν independent variables z_{kl} ($1 \leq k \leq l \leq n$) is called an automorphic form with the group Δ , if it is regular in E and satisfies there the equations

$$(4) \quad f(\mathfrak{Z}^*) = v(\mathfrak{A}, \mathfrak{B}) |\mathfrak{B}\mathfrak{Z} + \mathfrak{A}|^{-\nu} f(\mathfrak{Z})$$

for all transformations (1) in the group Δ , where g is a constant and the numbers $\nu = \nu(\mathfrak{A}, \mathfrak{B})$ depend only upon \mathfrak{A} and \mathfrak{B} . Let $L = L(\Delta, g, \nu)$ denote the set of all such functions $f(\mathfrak{Z})$, the weight g and the multiplier system ν being given. If f_1 and f_2 belong to this set, then so does $\lambda f_1 + \mu f_2$, for arbitrary complex constants λ and μ ; hence L is a vector space with a certain (finite or infinite) dimension d .

For automorphic forms of a single variable, i.e. in the case $n = 1$, the number d is given by the generalized Riemann-Roch theorem.⁴ It is not known in which way this theorem might be extended to automorphic forms of several variables. We now assume that the weight g is real and that all multipliers $\nu(\mathfrak{A}, \mathfrak{B})$ have absolute value 1. We shall derive a finite upper bound of d depending only upon n , g and δ .

Consider first the case $g = 0$. Then, by (4) the absolute value $\text{abs } f(\mathfrak{Z})$ is invariant under Δ ; consequently it attains in E a maximum at an inner point. This proves $f(\mathfrak{Z})$ is a constant, whence $d = 1$, if $\nu(\mathfrak{A}, \mathfrak{B}) = 1$, and $d = 0$ otherwise. In the remainder of the paper, we suppose $g \neq 0$.

LEMMA: Let $f(\mathfrak{Z})$ be a function of the set $L(\Delta, g, \nu)$, not identically 0. If all its partial derivatives of the orders $0, 1, \dots, h-1$ ($h \geq 0$) vanish at a point \mathfrak{Z}_0 of E , then $h \leq bg$.

⁴ E. Ritter, *Die multiplikativen Formen auf algebraischem Gebilde beliebigen Geschlechtes mit Anwendung auf die Theorie der automorphen Formen*, Math. Ann. 44, pp. 261-374 (1894).

H. Petersson, *Zur analytischen Theorie der Grenzkreisgruppen, Teil II*, Math. Ann. 115, pp. 175-204 (1938).

PROOF: The continuous function

$$\varphi(\mathfrak{Z}) = |\mathfrak{E} - \mathfrak{Z}\bar{\mathfrak{Z}}|^{1\theta} \text{ abs } f(\mathfrak{Z})$$

is invariant under Δ ; consequently it has in E a maximum $\mu > 0$, which is attained at a point \mathfrak{Z}_1 of F . On account of (4), we may assume that \mathfrak{Z}_0 also lies in F . In case $h > 0$, the function $f(\mathfrak{Z})$ vanishes at $\mathfrak{Z} = \mathfrak{Z}_0$, whence $\mathfrak{Z}_1 \neq \mathfrak{Z}_0$. In case $h = 0$, the assumption of the lemma holds for every point \mathfrak{Z}_0 of E , and we may suppose $\mathfrak{Z}_1 \neq \mathfrak{Z}_0$.

If the transformation (1) is any given element M of the group Ω , then the function $|\mathfrak{B}\mathfrak{Z} + \mathfrak{A}|^{-\theta} f(\mathfrak{Z}^*)$ belongs to $L(M^{-1}\Delta M, g, v)$. Since Ω is transitive in E and the diameter δ is invariant under Ω , we may assume for the proof of the lemma that $\mathfrak{Z}_0 = 0$ and $\rho(\mathfrak{Z}_1, 0) \leq \delta$. Let $\lambda_1, \dots, \lambda_n$ be the characteristic roots of $\mathfrak{Z}_1\bar{\mathfrak{Z}}_1$, $0 \leq \lambda_1 \leq \dots \leq \lambda_n$; then $0 < \lambda_n < 1$ and, by (2) and (3),

$$(5) \quad 0 < \sum_{k=1}^n \frac{\lambda_k}{1 - \lambda_k} \leq b.$$

We introduce a single complex variable z and choose in particular $\mathfrak{Z} = z\mathfrak{Z}_1$. For all points z of the circle $z\bar{z} < \lambda_n^{-1}$, the matrix \mathfrak{Z} lies in E ; hence there $f(\mathfrak{Z})$ is a regular analytic function $\psi(z)$ which vanishes at the point $z = 0$ at least of the order h and satisfies the relationship

$$\text{abs } \psi(z) = |\mathfrak{E} - z\bar{z}\mathfrak{Z}_1\bar{\mathfrak{Z}}_1|^{-1\theta} \varphi(z\mathfrak{Z}_1) \leq |\mathfrak{E} - z\bar{z}\mathfrak{Z}_1\bar{\mathfrak{Z}}_1|^{-1\theta} \mu,$$

where the equality holds for $z = 1$.

Let $1 < t < \lambda_n^{-1}$. On the circle $z\bar{z} \leq t$, the analytic function $z^{-h}\psi(z)$ attains the maximum of its absolute value at a point of the boundary, whence

$$\text{abs } \psi(1) \leq t^{-h} \max_{z\bar{z}=t} \text{abs } \psi(z)$$

$$|\mathfrak{E} - \mathfrak{Z}_1\bar{\mathfrak{Z}}_1|^{-1\theta} \mu \leq t^{-h} |\mathfrak{E} - t\mathfrak{Z}_1\bar{\mathfrak{Z}}_1|^{-1\theta} \mu.$$

But $|\mathfrak{E} - t\mathfrak{Z}_1\bar{\mathfrak{Z}}_1| = \prod_{k=1}^n (1 - t\lambda_k)$ and therefore

$$h \leq g \log \prod_{k=1}^n \frac{1 - \lambda_k}{1 - t\lambda_k} / \log t \quad (1 < t < \lambda_n^{-1}).$$

Performing the passage to the limit $t \rightarrow 1$, we obtain the inequality

$$(6) \quad h \leq g \sum_{k=1}^n \frac{\lambda_k}{1 - \lambda_k}.$$

The assertion of the lemma follows from (5) and (6).

THEOREM: The dimension d of $L(\Delta, g, v)$ is 0 for $g < b^{-1}$ and $\leq cg'$ for $g > 0$.

PROOF: Assume $d > 0$ and choose in $L(\Delta, g, v)$ a function $f(\mathfrak{Z})$, which does not vanish identically. Applying the lemma with $h = 0$, we infer $0 \leq bg$. This proves the theorem in the case $g < 0$.

Now consider the case $g > 0$. If $f(\mathfrak{Z}) \neq 0$ everywhere in E , then $f^{-1}(\mathfrak{Z})$ is a non-vanishing function of the set $L(\Delta, -g, v^{-1})$ and $-g < 0$, which is impossible. Consequently we may apply the lemma with $h = 1$ and obtain the

inequality $1 \leq bg$, whence $1 < (\nu + 1)(bg)^r = cg^r$. This proves the theorem in the case $g > 0$ and $d = 0$ or 1 .

In the remaining case $g > 0$, $d \geq 2$, let f_1, \dots, f_m be a finite number of linearly independent functions in $L(\Delta, g, \nu)$ and $m \geq 2$. We determine the positive integer h by the condition

$$(7) \quad \binom{\nu + h - 1}{\nu} < m \leq \binom{\nu + h}{\nu}$$

and choose m constants a_1, \dots, a_m , not all 0, such that all partial derivatives of the orders $0, 1, \dots, h - 1$ vanish for the function

$$f(\mathfrak{z}) = a_1 f_1 + \dots + a_m f_m$$

at the point $\mathfrak{z} = 0$; this is possible, by (7), since we have to satisfy $\binom{\nu + h - 1}{\nu}$ homogeneous linear equations with the m unknown quantities a_1, \dots, a_m . By (7) and the lemma,

$$m \leq \binom{\nu + h}{\nu} \leq (\nu + 1)h^r \leq (\nu + 1)(bg)^r = cg^r.$$

This proves the remaining part of the theorem.

3

A function $\chi(\mathfrak{z})$ is called an automorphic function with the group Δ , if $\chi(\mathfrak{z}) = f_1/f_0$, f_0 not identically 0, where f_1 and f_0 are automorphic forms in the same set $L(\Delta, g, \nu)$. For a sufficiently large value $G > 0$, certain functions in the set $L(\Delta, G, 1)$ can be expressed as Poincaré series,⁵ and it may be proved by known methods that there exist $\nu + 1$ of those functions, say F_0, \dots, F_ν , which are algebraically independent. Then the ν quotients $\chi_k = F_k/F_0$ ($k = 1, \dots, \nu$) are algebraically independent automorphic functions with the group Δ .

Define $q = [c\nu!G^r]$ and choose a positive integer Q satisfying the condition $q + 1 > c\nu!(G + gqQ^{-1})^r$. The number of power products

$$P = \chi^r \prod_{k=1}^{\nu} \chi_k^{s_k}$$

with $0 \leq r \leq q$, $0 \leq s_k$ ($k = 1, \dots, \nu$), $s_1 + \dots + s_\nu \leq Q$ is

$$(8) \quad A = (q + 1) \binom{Q + \nu}{\nu} > \frac{q + 1}{\nu!} Q^r > c(gq + GQ)^r;$$

we denote them by P_1, \dots, P_A . Then the A functions $f_0^q F_0^Q P_l$ ($l = 1, \dots, A$) are automorphic forms of the set $L(\Delta, gq + GQ, \nu^q)$; by (8) and the theorem, they are linearly dependent. Consequently, the automorphic function χ satisfies an algebraic equation of degree q whose coefficients are polynomials in χ_1, \dots, χ_ν and not all identically 0. Since q is fixed, the automorphic functions with the group Δ form an algebraic field with exactly ν independent elements.

INSTITUTE FOR ADVANCED STUDY

⁵ M. Sugawara, *Über eine allgemeine Theorie der Fuchschen Gruppen und Theta-Reihen*, Ann. of Math. (2) 41, pp. 488–494 (1940).

ON THE DERIVATIVES OF THE SECTIONS OF BOUNDED POWER SERIES¹

BY RHODA MANNING

(Received January 8, 1942)

1. Introduction

Let $f(z)$ represent a power series convergent in the open unit circle $|z| < 1$ and satisfying the condition $|f(z)| \leq 1$ in $|z| < 1$. It is well known that the sections $s_n(z)$ of $f(z)$ are not in general bounded in the open unit circle $|z| < 1$.² In 1925 L. Fejér³ proved that the sections $s_n(z)$, for all such functions $f(z)$, satisfy the condition $|s_n(z)| \leq 1$ in the circle $|z| \leq \frac{1}{2}$ for all n , and that this number $\frac{1}{2}$ cannot in general be replaced by a larger number.

Let r_n denote the radius of the largest circle $|z| \leq r_n$ in which the sections $s_n(z)$, for all functions $f(z)$ of the above type, satisfy the condition $|s_n(z)| \leq 1$. I. Schur and G. Szegő,⁴ extending Fejér's result, proved that the radii r_n constitute a monotone increasing sequence of algebraic numbers having the limit unity. They also studied the subclass of all functions $f(z)$ satisfying the additional condition $f(0) = 0$, and showed that, for all such functions $f(z)$, the radius R_n of the largest circle $|z| \leq R_n$ in which the condition $|s'_{n+1}(z)| \leq 1$ holds, for odd n , $n \geq 1$, satisfies the algebraic equation

$$1 - 2r - r^2 - (2n + 4)r^{n+1} - (2n + 2)r^{n+2} = 0.^5$$

Hence the sequence $\{R_n\}$, n odd, is ever increasing. The object of this note is to discuss the determination of the radii R_n in the case when n is even. The author has found in this case that the R_n , provided $n \geq 12$, satisfy the similar equation

$$1 - 2r - r^2 + (2n + 4)r^{n+1} + (2n + 2)r^{n+2} = 0.$$

Hence for even n , $n \geq 12$, the sequence $\{R_n\}$ is ever decreasing. Both sequences have the common limit $\rho = 2^{\frac{1}{2}} - 1$, the only positive root of the equation $1 - 2r - r^2 = 0$.

¹ Presented to the Society, December 2, 1939.

² L. Fejér, *Über gewisse Potenzreihen an der Konvergenzgrenze*, Sitzungsber. der math.-physik. Klasse der Bayer. Akad. der Wiss., 1910, Nr. 3.

³ L. Fejér, *Über die Positivität von Summen, die nach trigonometrischen oder Legendreschen Funktionen fortschreiten (Erste Mitteilung)*, Acta litt. ac sci. regiae univ. hung. Franciscose Josephinae, sectio sci. math., vol. 2, 1925, pp. 75-86.

⁴ I. Schur and G. Szegő, *Über die Abschnitte einer im Einheitskreise beschränkten Potenzreihe*, Sitzungsber. der Preuss. Akad. der Wiss., physik.-math. Klasse, 1925, in particular pp. 545-555.

⁵ Loc. cit. (4), p. 560.

It follows that the sections of $f'(z)$, for even n , $n \geq 12$, in general remain bounded by unity "longer" than the function $f'(z)$ itself, an unusual occurrence in this type of problem. As another immediate consequence of the theorem, we regain the well known fact that the derivative $f'(\frac{1}{2})$ cannot exceed unity in absolute value in the circle $|z| \leq 2^{\frac{1}{2}} - 1$, and that the bound $2^{\frac{1}{2}} - 1$ cannot in general be replaced by a larger one.⁶

In a thesis submitted to Stanford University⁷ the author has shown, by treating each case separately, that the numbers R_n , for the values of n excluded by the theorem, are also algebraic, and that they satisfy the following order relations:

$$R_1 < R_2 < R_3 < R_4 < R_5 < R_6 < R_7 < R_9 < R_8 < R_{11} < R_{13} < \dots \\ \dots < \rho = 2^{\frac{1}{2}} - 1 < \dots < R_{14} < R_{12} < R_{10}.$$

To facilitate computation of the radii R_n , $n \geq 12$, an asymptotic expression for R_n is given, of the form

$$R_n = \rho + (-1)^n a_n \rho^{n+1} + b_n \rho^{2n+1} + (-1)^n c'_n \rho^{3n+1},$$

where

$$a_n = \left(n + 1 + \frac{2^{\frac{1}{2}}}{2}\right), \quad b_n = (n+1)(n+2)a_n - \frac{1}{4}(2 - 2^{\frac{1}{2}})a_n^2,$$

and $0 < c'_n < 2a_n(n+1)^2(n+2)^2$.

Finally, a closely related theorem, stated by I. Schur and G. Szegő for odd values of n , is generalized to include large even values of n .⁸

2. Main theorem

Let $f(z)$ represent a power series convergent in the open unit circle $|z| < 1$ and satisfying the conditions $|f(z)| \leq 1$ in $|z| < 1$ and $f(0) = 0$. Let R_n denote the radius of the largest circle $|z| \leq R_n$ in which the section $s'_{n+1}(z)$, for all power series $f(z)$, satisfies the condition $|s'_{n+1}(z)| \leq 1$. If $n \geq 1$, $n \neq 2, 4, 6, 8, 10$, n an integer, then the radius R_n is the smallest positive root of the algebraic equation

$$G_n(r) = 1 - 2r - r^2 + (-1)^n[(2n+4)r^{n+1} + (2n+2)r^{n+2}] = 0.$$

PROOF. It has been shown⁹ that the radius of the largest circle $|z| \leq R_n$ in which the condition $|s'_{n+1}(z)| \leq 1$ holds, is the maximum value of r for which the harmonic polynomial

$$T_n(r, \phi) = \frac{1}{2} + 2r \cos \phi + 3r^2 \cos 2\phi + \dots + (n+1)r^n \cos n\phi$$

⁶ J. Dieudonné, *Polynomes et fonctions bornées d'une variable complexe*, École Normale Supérieure, Annales Scientifiques, vol. 48, 1930-31, p. 352.

⁷ Dissertation, Stanford University, June 1941.

⁸ Loc. cit. (4), pp. 558-559.

⁹ Loc. cit. (5).

remains non-negative, for all real values of ϕ . Let us denote the product

$$\begin{aligned} 2(1 - 2r \cos \phi + r^2)^2 \cdot T_n(r, \phi) &= 1 - 4r^2 + 4r^3 \cos \phi - r^4 \\ &- (2n + 4)r^{n+1} \cos (n + 1)\phi + 2r^{n+2}[(2n + 4) \cos n\phi + (n + 1) \cos (n + 2)\phi] \\ &- 2r^{n+3}[(n + 2) \cos (n - 1)\phi + (2n + 2) \cos (n + 1)\phi] + (2n + 2)r^{n+4} \\ &\quad \cdot \cos n\phi \end{aligned}$$

by $F_n(r, \phi)$. Since the cosine is an even function, we need only consider values of ϕ satisfying $0 \leq \phi \leq \pi$.

If n is odd, then $F_n(r, \pi)$ is an obvious lower boundary for $F_n(r, \phi)$. Since further $F_n(r, \pi) = (1 + r)^2 \cdot G_n(r)$, R_n is the only positive root of the equation $G_n(r) = 0$.

Now let n be even. If $r = 0.42$ and $n \geq 8$, then

$$\begin{aligned} F_n(r, \pi) &= (1 + r)^2(1 - 2r - r^2 + (2n + 4)r^{n+1} + (2n + 2)r^{n+2}) \\ &< (1 + r)^2(-0.0164 + 0.0113) < 0, \end{aligned}$$

whence $R_n < 0.42$, $n \geq 8$. Hence we may restrict our proof of the inequality $F_n(r, \phi) \geq F_n(r, \pi)$, $n \geq 12$, to values of r contained in the interval $0 < r < 0.42$.

Now

$$\begin{aligned} F_n(r, \phi) - F_n(r, \pi) &= 4r^3(1 + \cos \phi) - 2r^{n+1}[(n + 2)(1 + \cos (n + 1)\phi) \\ &\quad + \{(2n + 4)(1 - \cos n\phi) + (n + 1)(1 - \cos (n + 2)\phi)\}r \\ &\quad + \{(n + 2)(1 + \cos (n - 1)\phi) + (2n + 2)(1 + \cos (n + 1)\phi)\}r^2 \\ &\quad + (n + 1)(1 - \cos n\phi)r^3]. \end{aligned}$$

On dividing this difference by the positive quantity $2r^3(1 + \cos \phi)$, we notice that all the terms except the first in the resulting expression are of the form

$$-c \frac{1 + (-1)^{k-1} \cos k\phi}{1 + \cos \phi} = -c \frac{1 - \cos k(\pi - \phi)}{1 - \cos (\pi - \phi)},$$

$c > 0$, $k = 1, 2, 3, \dots$. It is easily verified that this expression is never less than $-ck^2$, and that it attains this value for $\phi = \pi$. Hence the inequality to be proved is equivalent to

$$\begin{aligned} \lim_{\phi \rightarrow \pi} \frac{F_n(r, \phi) - F_n(r, \pi)}{2r^3(1 + \cos \phi)} &= 2 - r^{n-2}[(n + 2)(n + 1)^2 \\ &\quad + \{(2n + 4)n^2 + (n + 1)(n + 2)^2\}r + \{(n + 2)(n - 1)^2 \\ &\quad + (2n + 2)(n + 1)^2\}r^2 + (n + 1)n^2r^3] \geq 0, \end{aligned}$$

which holds since we can satisfy simultaneously

$$r^{n-2}(n + 2)^3(1 + r)^3 < 2, \quad 0 < r < 0.42 \quad \text{and} \quad n \geq 12.$$

Hence for even n , $n \geq 12$, and $0 < r < 0.42$,

$$F_n(r, \phi) \geq F_n(r, \pi) = (1 + r)^2 \cdot G_n(r),$$

whence R_n is the smallest positive root of the algebraic equation $G_n(r) = 0$.

3. Asymptotic Inequalities

It will be assumed in the two cases which follow that $n \geq 12$.

CASE 1. n odd. The following derivation of an upper bound for R_n depends on the remark that R_n is the only positive root of the equation $G_n(r) = 0$. Since $G_n(0) = 1 > 0$ and $G_n(r)$ is ever decreasing for positive values of r , we conclude that if $G_n(r) < 0$ for $r = r_0$, say, then $R_n < r_0$.

We shall suppose that $r = \rho(1 - x)$, where $x = a_n \rho^n - b_n \rho^{2n}$, $(n + 1 + \frac{1}{2}2^{\frac{1}{2}}) = a_n$, $b_n = (n + 1)(n + 2)a_n - \frac{1}{4}(2 - 2^{\frac{1}{2}})a_n^2$, and show that with this choice of r , $G_n(r) < 0$. We note that $0 < x < a_n \rho^n < 0.0005$, since $\rho^{12} = 0.0000255 \dots$ and $n^3 \rho^n$ is a decreasing function of n . Therefore $r > 0$ and for $k = 1, 2$, $r^{n+k} > \rho^{n+k}[1 - (n + k)x]$.¹⁰ Hence

$$\begin{aligned} G_n(r) &< 1 - 2r - r^2 - (2n + 4)\rho^{n+1}[1 - (n + 1)x] - (2n + 2)\rho^{n+2}[1 - (n + 2)x] \\ &= 2(2^{\frac{1}{2}})\rho x - \rho^2 x^2 - 2(2^{\frac{1}{2}})\rho^{n+1}[a_n - (n + 1)(n + 2)x] \\ &= 2(2^{\frac{1}{2}})\rho(a_n \rho^n - b_n \rho^{2n}) - \rho^2(a_n^2 \rho^{2n} - 2a_n b_n \rho^{3n} + b_n^2 \rho^{4n}) \\ &\quad - 2(2^{\frac{1}{2}})\rho^{n+1}[a_n - (n + 1)(n + 2)a_n \rho^n + (n + 1)(n + 2)b_n \rho^{2n}] \\ &= -b_n \rho^{3n+1}[2(2^{\frac{1}{2}})(n + 1)(n + 2) - 2a_n \rho + b_n \rho^{n+1}] < 0. \end{aligned}$$

To derive a lower bound for R_n , we notice that if $G_n(r) > 0$, then $r < R_n$. Set $r = \rho(1 - x)$, where $x = a_n \rho^n - b_n \rho^{2n} + c_n \rho^{3n}$, with a_n, b_n as before, and $c_n = 2a_n(n + 1)^2(n + 2)^2$. On expanding $(1 - x)^{n+k}$ as an exponential series with $(n + k) \log(1 - x)$ as argument, we find $r^{n+k} < \rho^{n+k}[1 - (n + k)x + n^2 x^2]$, $k = 1, 2$. Hence

$$\begin{aligned} G_n(r) &> 2(2^{\frac{1}{2}})\rho x - \rho^2 x^2 - 2(2^{\frac{1}{2}})\rho^{n+1}[a_n - (n + 1)(n + 2)x + a_n n^2 x^2] \\ &> 2(2^{\frac{1}{2}})\rho^{n+1}(a_n - b_n \rho^n + c_n \rho^{2n}) - a_n^2 \rho^{2n+2} - 2(2^{\frac{1}{2}})\rho^{n+1}[a_n \\ &\quad - (n + 1)(n + 2)\rho^n(a_n - b_n \rho^n + c_n \rho^{2n}) + n^2 a_n^2 \rho^{2n}] \\ &= 2(2^{\frac{1}{2}})\rho^{3n+1}[c_n - (n + 1)(n + 2)b_n + (n + 1)(n + 2)c_n \rho^n - n^2 a_n^2] > 0, \end{aligned}$$

since $x < a_n \rho^n$ and $c_n > 2(n + 1)(n + 2)b_n > 2n^2 a_n^2$.

CASE 2. n even. To derive a lower bound for R_n , we notice that $G_n(r) = 0$ has two positive roots, the smaller of which is R_n , and that $G_n(0) = 1 > 0$. Hence if $G_n(r) > 0$ and $G'_n(r) < 0$ simultaneously, then $r < R_n$. Let $r = \rho(1 + x)$, where $x = a_n \rho^n + b_n \rho^{2n}$. A repetition of the argument in the first part of Case 1 gives $r^{n+k} > \rho^{n+k}[1 + (n + k)x]$, $k = 1, 2$, and that $G_n(r) > 0$. That $G'_n(r) < 0$, for $0 < r < \frac{1}{2}$, is trivial.

¹⁰ Hardy, Littlewood and Polya, *Inequalities*, p. 40.

To find an upper bound for R_n , we note that if $G_n(r) < 0$, then $R_n < r$. Set $r = \rho(1+x)$, where $x = a_n\rho^n + b_n\rho^{2n} + c_n\rho^{3n}$. Then $x < (n+2)\rho^n$, and with a little attention it can be seen that $r^{n+k} = \exp[(n+k) \log \rho(1+x)] < \rho^{n+k}[1 + (n+k)x + (n+1)^2x^2]$, $k = 1, 2$. Hence

$$\begin{aligned} G_n(r) &< -2(2^{\frac{1}{2}})\rho x - \rho^2 x^2 + 2(2^{\frac{1}{2}})\rho^{n+1}[a_n + (n+1)(n+2)x + a_n(n+1)^2x^2] \\ &< -2(2^{\frac{1}{2}})\rho^{n+1}(a_n + b_n\rho^n + c_n\rho^{2n}) - a_n^2\rho^{2n+2} - 2a_nb_n\rho^{3n+2} \\ &\quad + 2(2^{\frac{1}{2}})\rho^{n+1}[a_n + (n+1)(n+2)\rho^n(a_n + b_n\rho^n + c_n\rho^{2n}) + a_n(n+1)^2 \\ &\quad \quad \quad (n+2)^2\rho^{2n}] \\ &= - (2^{\frac{1}{2}})\rho^{3n+1}[c_n - 2b_n(n+1)(n+2) - 2c_n(n+1)(n+2)\rho^n + \\ &\quad \quad \quad (2^{\frac{1}{2}})a_nb_n\rho] < 0. \end{aligned}$$

4. A Related Theorem

The method of proof of the main theorem applies to the following theorem:

Let $f(z)$ represent a power series convergent in the open unit circle $|z| < 1$ and satisfying the condition $|f(z)| \leq 1$ in $|z| < 1$. Let $\alpha < 0, \beta > 0, \alpha + \beta = 1$, and let r_n be the radius of the largest circle $|z| \leq r_n$ in which the sections $s_n(z)$, for all power series $f(z)$, satisfy the condition $|\alpha s_0(z) + \beta s_n(z)| \leq 1$. Then for odd $n, n \geq 1$, and for sufficiently large even n , the radii r_n satisfy the algebraic equation

$$1 + (\alpha - \beta)r + (-1)^n 2\beta r^{n+1} = 0.$$

Hence $\lim_{n \rightarrow \infty} r_n = \frac{1}{\beta - \alpha}.$

PROOF. The radius r_n is the maximum value of r for which the cosine polynomial

$$T_n(r, \phi) = \alpha/2 + \beta(\frac{1}{2} + r \cos \phi + r^2 \cos 2\phi + \dots + r^n \cos n\phi)$$

remains non-negative, for all real values of ϕ .¹¹ Let

$$\begin{aligned} F_n(r, \phi) &= 2(1 - 2r \cos \phi + r^2) \cdot T_n(r, \phi) \\ &= \alpha(1 - 2r \cos \phi + r^2) + \beta[1 - r^2 + 2r^{n+2} \cos n\phi - 2r^{n+1} \cos (n+1)\phi]. \end{aligned}$$

If $r \geq 1$,

$$\begin{aligned} F_n\left(r, \frac{\pi}{n}\right) &\leq \alpha(1 - r)^2 + \beta\left[1 - r^2 - 2r^{n+1}\left(r + \cos \frac{n+1}{n}\pi\right)\right] \\ &\leq -2\beta\left(1 + \cos \frac{n+1}{n}\pi\right) < 0. \end{aligned}$$

Hence $r_n < 1$ for all n .

¹¹ Loc. cit. (4), p. 558.

Now

$$\begin{aligned} F_n(r, \pi) &= \alpha(1 + 2r + r^2) + \beta(1 - r^2 + (-1)^n 2r^{n+1} + (-1)^n 2r^{n+2}) \\ &= 1 + 2\alpha r + (\alpha - \beta)r^2 + (-1)^n 2\beta r^{n+1} + (-1)^n 2\beta r^{n+2} \\ &= (1 + r)(1 + (\alpha - \beta)r + (-1)^n 2\beta r^{n+1}). \end{aligned}$$

If n is odd, then $F_n(r, \pi)$ is an obvious lower boundary for $F_n(r, \phi)$, for all real ϕ .

Let n be even. We shall show that the inequality

$$F_n(r, \phi) \geq F_n(r, \pi), \quad r = r_n,$$

holds for all real ϕ , and for sufficiently large even n .

Rewritten, it assumes the form

$$-2\alpha r(1 + \cos \phi) - 2\beta r^{n+1}[(1 + \cos(n+1)\phi + (1 - \cos n\phi)r] \geq 0.$$

The substitution $\phi = \pi - x$ yields

$$-\alpha(1 - \cos x) - \beta r^n[(1 - \cos(n+1)x + (1 - \cos nx)r] \geq 0,$$

whence, dividing by the positive quantity $(1 - \cos x)$, and taking the limit as $x \rightarrow 0$, we obtain the inequality

$$-\alpha \geq \beta r^n[(n+1)^2 + n^2 r],$$

which holds for sufficiently large n and $r < 1$. But $r_n < 1$ for all n . Hence r_n is a root of the algebraic equation

$$1 + (\alpha - \beta)r + 2\beta r^{n+1} = 0.$$

OREGON STATE COLLEGE,
CORVALLIS, ORE.

THE TRANSFORMATION T OF CONGRUENCES

By V. G. GROVE

(Received June 11, 1941)

1. Introduction

We propose to study in this paper a certain relationship between congruences in a projective space of three dimensions. Analytical conditions for this relationship, called by us the transformation T , were developed by Cook¹ in a form slightly different from that used in this paper. Fubini² also called attention to this relationship somewhat earlier; but neither of these papers showed the relationship between the transformation T and the theory of W -congruences. The present paper is more closely allied with a recent paper by Fubini³ on W -congruences.

Two congruences Γ and $\bar{\Gamma}$ will be said to be in the *relation of a transformation T* , if the lines of the congruences are in one-to-one correspondence such that

1. corresponding lines are not coplanar,
2. the developables of the congruences correspond, and
3. such that there exists at least three transversal surfaces⁴ of each congruence whose tangent planes at their points of intersection with the line of that congruence pass through the corresponding line of the other congruence.

The transformation T is of two types, one of which we have called the asymptotic type, and the other the conjugate type. Associated with each of these types there is a one-parameter family, or pencil of congruences. This pencil seems to be somewhat more general than the pencil⁵ defined by Fubini. As in the case of Fubini's pencils, we find that if one congruence of the associated pencil is a W -congruence, all congruences of the pencil are W . Associated with the transformation T are four congruences such that if any three are W -congruences, the other is also.

Let the curves which correspond to the developables of Γ and $\bar{\Gamma}$ be chosen as the parametric curves on the focal surfaces S_* , S_w , S_* , S_w of Γ and $\bar{\Gamma}$. Then the homogeneous projective coordinates z_i , w_i , \bar{z}_i , \bar{w}_i , $i = 1, 2, 3, 4$, of the focal points on the lines of the congruences satisfy differential equations of

¹ A. J. Cook, *Pairs of rectilinear congruences with generators in one-to-one correspondence*, Trans. Am. Math. Soc., Vol. 32 (1930), pp. 31-46.

² G. Fubini, *Su alcune classi di congruenze di rette e sulle trasformazioni delle Superficie R*, Annali di Matematica, (4), Vol. 1 (1923-24), pp. 241-257.

³ G. Fubini, *On Bianchi's permutability theorem and the theory of W -congruences*, these Annals, Vol. 41 (1940), pp. 620-638.

⁴ A. J. Cook, loc. cit., says that each of the congruence has the intersector property I with respect to the other congruence.

⁵ G. Fubini, loc. cit., p. 634.

the form

$$\begin{aligned}
 (1) \quad z_u &= f\bar{z} + g\bar{w} + rz + sw, & \bar{z}_u &= \bar{f}z + \bar{g}w + \bar{r}\bar{z} + \bar{s}\bar{w}, \\
 z_v &= mz + nw, & \bar{z}_v &= \bar{m}\bar{z} + \bar{n}\bar{w}, \\
 w_u &= Nz + Mw, & \bar{w}_u &= \bar{N}\bar{z} + \bar{M}\bar{w}, \\
 w_v &= Gz + Fw + Sz + Rw, & \bar{w}_v &= \bar{G}\bar{z} + \bar{F}\bar{w} + \bar{S}\bar{z} + \bar{R}\bar{w}.
 \end{aligned}$$

The integrability conditions of system (1) may be written in the form

$$\begin{aligned}
 (2a) \quad m_u - r_v + nN - sS &= g\bar{G}, \\
 s_v - n_u + s(R - m) + n(r - M) &= -g\bar{F}, \\
 f_v + f(\bar{m} - m) + g\bar{S} + sG &= 0, \\
 g_v + g(\bar{R} - m) + sF + f\bar{n} &= 0;
 \end{aligned}$$

$$\begin{aligned}
 (2b) \quad M_v - R_u + nN - sS &= \bar{g}G, \\
 S_u - N_v + S(r - M) + N(R - m) &= -\bar{f}G, \\
 F_u + F(\bar{M} - M) + G\bar{s} + gS &= 0, \\
 G_u + G(\bar{r} - M) + fS + F\bar{N} &= 0,
 \end{aligned}$$

and two other sets obtained from these by placing bars above the letters where they do not occur and removing those which do occur.

Let us denote by x, y, z , etc. the points whose homogeneous projective coordinates are x_i, y_i, z_i , etc. ($i = 1, 2, 3, 4$). Let S_x be a transversal surface of $\bar{\Gamma}$ generated by the point x whose coordinates x_i are of the form $x = \bar{w} + \lambda\bar{z}$. It follows that

$$\begin{aligned}
 (3) \quad x_u &= (\bar{M} + \lambda\bar{s})x + \lambda(\bar{f}z + \bar{g}w) + L_1\bar{z}, \\
 x_v &= (\bar{R} + \lambda\bar{n})x + \bar{G}z + \bar{F}w + L_2\bar{z},
 \end{aligned}$$

wherein

$$\begin{aligned}
 L_1 &= \lambda_u - [\bar{s}\lambda^2 + (\bar{M} - \bar{r})\lambda - \bar{N}], \\
 L_2 &= \lambda_v - [\bar{n}\lambda^2 + (\bar{R} - \bar{m})\lambda - \bar{S}].
 \end{aligned}$$

The tangent plane to S_x at x passes through the line g of Γ if and only if $L_1 = L_2 = 0$. If one equates the derivatives λ_{uv} and λ_{vu} computed from $L_1 = 0$, and $L_2 = 0$, one finds, by using the integrability conditions (2), that these latter two equations can have analytic solutions only when the equation

$$(4) \quad g\bar{F}\lambda^2 - (\bar{g}G + g\bar{G})\lambda + \bar{f}G = 0$$

is satisfied. It follows from the third property demanded of two congruences that they be in the relation of a transformation T , that the coefficients of (4) vanish. Hence

$$(5) \quad g\bar{F} = \bar{g}G + g\bar{G} = \bar{f}G = 0.$$

In a similar manner one may show that

$$(6) \quad \bar{g}F = \bar{g}G + g\bar{G} = f\bar{G} = 0.$$

Hence the congruences Γ and $\bar{\Gamma}$ are in relation T if and only if conditions (5) and (6) are satisfied.

In general the tangent planes at z and w to S_z , S_w do not coincide. Hence

$$(7) \quad f\bar{F} - \bar{g}\bar{G} \neq 0, \quad fF - gG \neq 0.$$

From (5), (6), (7) we find that the transformation T is of two types; we call these types respectively

(i) the *asymptotic type* if

$$(8) \quad f = \bar{f} = F = \bar{F} = g\bar{G} + \bar{g}G = 0;$$

and the

(ii) *conjugate type* if

$$(9) \quad g = \bar{g} = G = \bar{G} = 0.$$

2. The Asymptotic Type

Let us consider first the asymptotic type of the transformation T . Under the conditions (8), we observe first that the integrability conditions (2) imply that $s = S = \bar{s} = \bar{S} = 0$.

Let λ_1, λ_2 be two distinct solutions of $L_1 = L_2 = 0$, and suppose that these solutions determine the two transversal surfaces S_z, S_w of $\bar{\Gamma}$. Then from (3) we observe that x, y, z and w satisfy equations of the form

$$(10) \quad \begin{aligned} x_u &= ax + bw, & y_u &= a'y + b'w, \\ x_v &= Ax + Bz, & y_v &= A'y + B'z, \\ z_u &= px + qy + rz, & w_u &= Nz + Mw, \\ z_v &= mz + nw, & w_v &= Qx + Py + Rw, \end{aligned}$$

with

$$(11) \quad bp + b'q = 0, \quad BQ + B'P = 0, \quad bb'BB' \neq 0.$$

The integrability conditions of system (10) are

$$(12) \quad \begin{aligned} a_v + bQ &= A_u + Bp, & A'_u + B'q &= a'_v + b'P, \\ bP &= Bq, & B'p &= b'Q, \\ B_u + Br &= aB, & b'_v + b'R &= A'b', \\ b_v + bR &= Ab, & B'_u + B'r &= a'B'. \end{aligned}$$

$$(13) \quad \begin{aligned} p_v + Ap &= mp, & P_u + a'P &= MP, \\ q_v + A'q &= mq, & Q_u + aQ &= MQ, \\ r_v + Bp + B'q &= nN + m_u, & R_u + b'P + bQ &= nN + M_v, \\ n_u + Mn &= nr, & N_v + mN &= NR. \end{aligned}$$

From the first of (2a) and (2b) and the third of (13) we find that $g\bar{G} + \bar{g}G = 0$ implies that

$$pB + qB' + b'P + bQ = 0.$$

From (12) and (13) we see that

$$\begin{aligned} a_v - a'_v &= A_u - A'_u, \\ r_v + M_v + a_v + a'_v &= R_u + m_u + A_u + A'_u, \end{aligned}$$

and we may verify that, by a transformation of the form

$$(14) \quad x = \lambda x', \quad y = \mu y', \quad z = \nu z', \quad w = \rho w',$$

we may make

$$(15) \quad \begin{aligned} a &= a', \quad A = A', \\ r + M + 2a &= R + m + 2A = 0. \end{aligned}$$

We shall assume that this transformation has been effected. The conditions (15) are maintained by transformations (14) with

$$\lambda/\mu = \text{const.} \quad \lambda\mu\nu\rho = \text{const.}$$

Again from (12) we note that

$$(16) \quad \frac{\partial}{\partial u} \log \frac{B}{B'} = 0, \quad \frac{\partial}{\partial v} \log \frac{b}{b'} = 0.$$

And hence from (11) and (16) we may write

$$(17) \quad q = pU, \quad Q = PV, \quad b = -b'\dot{U}, \quad B' = -BV$$

wherein U and V are respectively functions of u and v alone.

The focal points \bar{z} , \bar{w} of \bar{g} are readily found to be determined by the formulas

$$(18) \quad \bar{z} = (B'x - By)/D, \quad \bar{w} = (by - b'x)/D, \quad D = bB' - b'B.$$

Hence we can recover equations (1) in the form

$$(19) \quad \begin{aligned} z_u &= pB(1 - UV)\bar{w} + rz, & \bar{z}_u &= w + \bar{r}\bar{z}, \\ z_v &= mz + nw, & \bar{z}_v &= \bar{m}\bar{z} + \bar{n}\bar{w}, \\ w_u &= Nz + Mw, & \bar{w}_u &= \bar{N}\bar{z} + \bar{M}\bar{w}, \\ w_v &= b'P(1 - UV)\bar{z} + Rw, & \bar{w}_v &= \bar{z} + \bar{R}\bar{w}, \end{aligned}$$

wherein

$$(20) \quad \begin{aligned} \bar{r} &= 2a - r - (\log D)_u, & \bar{m} &= R, \\ \bar{R} &= 2A - R - (\log D)_v, & \bar{M} &= r, \\ \bar{n} &= \frac{BV_v}{b'(1 - UV)}, & \bar{N} &= \frac{b'U_u}{B(1 - UV)}, \\ D &= b'B(UV - 1), & \Delta &= pP(1 - UV). \end{aligned}$$

3. W -congruences in the Asymptotic Case

The differential equations of the asymptotic curves on S_z , S_w are readily found to be respectively

$$(21) \quad \begin{aligned} pU_u du^2 + nP(1 - UV) dv^2 &= 0, \\ Np(1 - UV) du^2 + PV_v dv^2 &= 0. \end{aligned}$$

Hence Γ is a W -congruence if and only if the invariant

$$W = nN - \frac{U_u V_v}{(1 - UV)^2}$$

vanishes.

The differential equations of the asymptotic curves on S_z , S_w are found to be

$$\begin{aligned} \bar{N} du^2 + n dv^2 &= 0, \\ N du^2 + \bar{n} dv^2 &= 0. \end{aligned}$$

It follows that $\bar{\Gamma}$ is a W -congruence if and only if the invariant

$$\bar{W} = \bar{n}\bar{N} - nN$$

vanishes. But from (20) we see that

$$(22) \quad W + \bar{W} = 0.$$

It follows *therefore that, if one of two congruences in the relation of a transformation T of the asymptotic type is a W -congruence, the other is also.*

4. The Transversal Surfaces

If from (10) one eliminates z and w , it will be found that the coordinates of the current points x , and y of S_z , S_y satisfy the equations

$$(23) \quad \begin{aligned} x_{uu} &= \theta x_u - \frac{b'NU}{B} x_v + ()x, \\ x_{uv} &= Ax_u + ax_v + ()x - b'PUy, \\ x_{vv} &= -\frac{Bn}{b'U} x_u + \varphi x_v + ()x, \\ y_{uu} &= \theta'y_u - \frac{b'N}{B'V} y_v + ()y, \\ y_{uv} &= Ay_u + ay_v + ()y - b'PVx, \\ y_{vv} &= -\frac{BnV}{b'} y_u + \varphi'y_v + ()y, \end{aligned}$$

wherein the omitted coefficients are immaterial for our purposes, and wherein

$$(25) \quad \begin{aligned} \theta &= a + M + (\log b'U)_u, & \theta' &= a + M + (\log b')_u, \\ \varphi &= A + m + (\log B)_v, & \varphi' &= A + m + (\log BV)_v. \end{aligned}$$

It follows from (23) and (24) that the curves on S_x, S_y which correspond to the developables of Γ and $\bar{\Gamma}$ form asymptotic nets N_x, N_y on those surfaces. Moreover those surfaces are not ruled surfaces.

Denote by I_x, I_y the invariants whose vanishing imply that N_x or N_y is isothermally asymptotic. We find that

$$I_x = I_y = \frac{\partial^2}{\partial u \partial v} \log \left(\frac{b'^2 N}{B^2 n} \right) = 4(A_u - a_v).$$

But the vanishing of this function, as is seen from the first of (12), implies that $pP - qQ$ vanishes. Hence neither N_x nor N_y is isothermally asymptotic.

Suppose S_ξ is a transversal surface of $\bar{\Gamma}$ distinct from S_x, S_y . If the coordinates of ξ are defined by

$$\xi = y + \lambda x$$

it follows that the tangent plane to S_ξ at ξ passes through g if and only if λ is a constant. We may say that the line (ξw) (or (ξ, z)) generates a pencil of congruences. They are the asymptotic tangents to the one focal surface S_ξ , the locus of ξ being the line \bar{g} . We note that $I_\xi = I_x = I_y$ for every λ .

If we denote the coordinates of the tangent planes to $S_x, S_y, S_\xi, S_{\bar{\xi}}$ respectively by ξ, η, ω, ζ we find that these functions satisfy the following system of differential equations

$$\begin{aligned} (26) \quad \xi_u &= -a\xi + q\omega, & \eta_u &= -a\eta + p\omega, \\ \xi_v &= -A\xi + P\zeta, & \eta_v &= -A\eta + Q\zeta, \\ \zeta_u &= b'\xi + b\eta - M\zeta, & \omega_u &= -N\zeta - r\omega, \\ \zeta_v &= -R\zeta - n\omega, & \omega_v &= B'\xi + B\eta - m\omega. \end{aligned}$$

We have said⁶ that two nets are in relation C if the developables of the congruence of lines joining corresponding points of the nets intersect the sustaining surfaces of the nets in those nets. In particular two conjugate nets in the relation of a fundamental transformation F are in relation C . Two nets in relation C are said to be K_α transforms if

$$\frac{\partial^2}{\partial u \partial v} \log \alpha = 0$$

where α is one of the cross ratios of the corresponding points of the nets and the two focal points on the line of the congruence through these points. In particular nets in the relation of a transformation of Koenig are K_α transforms.

We readily verify that $\alpha = bB'/(b'B)$. From (17) we note that $\alpha = UV$. Hence N_x, N_y are K_α transforms in the asymptotic case. Similarly from (26) we see that $N_\xi, N_{\bar{\xi}}$ are also K_α transforms since in that case $\alpha = UV$.

⁶ V. G. Grove, *Transformations of Nets*, Trans. Am. Math. Soc., Vol. 30 (1928), pp. 483-497. *Ibid.*, p. 493.

5. The Focal Surfaces

From (19) we see that the functions z and w satisfy equations of the form

$$\begin{aligned} z_{uv} &= mz_u + rz_v + (\)z, \\ w_{uv} &= Rw_u + Mw_v + (\)w, \end{aligned}$$

the omitted coefficients being immaterial for our purposes. The nets N_s , N_w have equal point invariants if the respective invariants E_s , E_w defined by

$$(27) \quad E_s = m_u - r_v, \quad E_w = M_v - R_u,$$

vanish.

It is readily seen that

$$2(E_w - E_s) = I_s = I_v.$$

Hence *not both N_s and N_w can have equal point invariants.*

From (21) we find that N_s and N_w are isothermally conjugate if the respective invariants

$$\begin{aligned} I_s &= \frac{\partial^2}{\partial u \partial v} \log \frac{pU_u}{nP(1 - UV)}, \\ I_w &= \frac{\partial^2}{\partial u \partial v} \log \frac{PV_v}{Np(1 - UV)}, \end{aligned} \quad (28)$$

vanish. But we may show that

$$(29) \quad I_w - I_s = I_s.$$

Hence *not both N_s , N_w can be isothermally conjugate.*

6. The Conjugate Type

The conjugate type of the transformation T is characterized by the conditions $g = \bar{g} = G = \bar{G} = 0$. Let S_x , S_y be two transversal surfaces of $\bar{\Gamma}$ whose tangent planes pass through the lines g of Γ . Then from (3) and by use of a transformation of the form (14) we may show that x , y , z , w may be made to satisfy equations of the form

$$\begin{aligned} x_u &= bz, & y_u &= b'z, \\ x_v &= Bw, & y_v &= B'w, \\ z_u &= px + qy - Mz + sw, & w_u &= Nz + Mw, \\ z_v &= mz + nw, & w_v &= Qx + Py + Sz - mw, \end{aligned} \quad (30)$$

wherein

$$(31) \quad Bp + B'q = bQ + b'P = 0, \quad m_u = M_v.$$

The integrability conditions of system (30) are

$$\begin{aligned}
 (32) \quad & b_v + bm = BN, & B'_u + B'M = b'n, \\
 & B_u + BM = bn, & b'_v + b'm = B'N, \\
 & P_u + qS = MP, & p_v + Qs = mp, \\
 & Q_u + pS = MQ, & q_v + Ps = mq, \\
 & sS - nN = 2m_u = 2M_v, \\
 & s_v - n_u = 2(ms + Mn), & S_u - N_v = 2(MS + mN).
 \end{aligned}$$

System (30) is preserved under all transformations of the form (14) with $\lambda = \text{const.}$, $\mu = \text{const.}$, $\rho\sigma = \text{const.}$

The focal points \bar{z} , \bar{w} on the line \bar{g} of $\bar{\Gamma}$ are determined by the formulas

$$(33) \quad \bar{z} = (B'x - By)/D, \quad \bar{w} = (by - b'x)/D, \quad D = bB' - b'B.$$

Hence we may recover equations (1) in the form

$$\begin{aligned}
 (34) \quad & z_u = f\bar{z} - Mz + sw, & \bar{z}_u = z - [M + (\log D)_u]\bar{z} - n\bar{w}, \\
 & z_v = mz + nw, & \bar{z}_v = m\bar{z} + \bar{n}\bar{w}, \\
 & w_u = Nz + Mw, & \bar{w}_u = \bar{N}\bar{z} + M\bar{w}, \\
 & w_v = F\bar{w} + Sz - mw, & \bar{w}_v = w - N\bar{z} - [m + (\log D)_v]\bar{w},
 \end{aligned}$$

wherein

$$\bar{n} = (BB'_v - B'B_v)/D, \quad \bar{N} = (b'b_u - bb'_u)/D.$$

7. W -Congruences in the Conjugate Case

Denote by Γ_{11} the congruence of lines $(x z)$, Γ_{22} the congruence of lines (y, w) , Γ_{21} that formed by $(y z)$, Γ_{12} that formed by $(x w)$. Let W_{ij} be an invariant whose vanishing implies that Γ_{ij} is a W -congruence. Four such functions are:

$$\begin{aligned}
 (35) \quad & W_{11} = \frac{\partial^2}{\partial u \partial v} \log \frac{q}{B} - 4m_u, & W_{12} = \frac{\partial^2}{\partial u \partial v} \log \frac{P}{b} - 4m_u, \\
 & W_{21} = \frac{\partial^2}{\partial u \partial v} \log \frac{p}{B'} - 4m_u, & W_{22} = \frac{\partial^2}{\partial u \partial v} \log \frac{Q}{b'} - 4m_u.
 \end{aligned}$$

The congruences Γ , $\bar{\Gamma}$ are W -congruences if the respective invariants

$$\begin{aligned}
 (36) \quad & W = \frac{\partial^2}{\partial u \partial v} \log \Delta - 4m_u, & \bar{W} = -\frac{\partial^2}{\partial u \partial v} \log D - 4m_u, \\
 & (\Delta = pP - qQ)
 \end{aligned}$$

vanish.

But using (31) we show easily that

$$(37) \quad \frac{\Delta}{D} = \frac{qQ}{b'B} = \frac{pP}{bB'}.$$

Hence

$$(38) \quad W + \bar{W} = W_{11} + W_{22} = W_{12} + W_{21}.$$

From the points x, y, z, w there may be formed three different skew quadrilaterals. From (38) we may say that *if any three of the sides of these quadrilaterals generate W -congruences so also does the fourth side.* If we agree to say that the tangents to a family of asymptotic curves on a surface form a W -congruence, we note that equation (22) is then a special case of (38).

Let S_ξ be a transversal surface of $\bar{\Gamma}$ whose tangent plane at ξ passes through the line g of Γ . Then if

$$(39) \quad \xi = x + \lambda y$$

it follows that $\lambda = \text{const.}$ We find readily that

$$(40) \quad \xi_u = (b + \lambda b')z, \quad \xi_v = (B + \lambda B')w,$$

and if $\Gamma_{11}^\lambda, \Gamma_{12}^\lambda$ are the congruences of lines (ξ, z) and (ξ, w) respectively and W_{ij}^λ the corresponding invariants W , then

$$(41) \quad W_{11}^\lambda = W_{11}, \quad W_{12}^\lambda = W_{12}.$$

We may say that the congruences Γ_{11}^λ and Γ_{12}^λ form pencils. We shall call them *the associated pencils*. It follows from (41) that *if one congruence of an associated pencil is a W -congruence, all congruences of the pencil are W -congruences.*

Denote by ρ_{ij} the focal points (other than x or y) on the lines of the congruences Γ_{ij} . We find that these points are defined by

$$(42) \quad \begin{aligned} \rho_{11} &= Bz - nx, & \rho_{12} &= bw - Nx, \\ \rho_{21} &= B'z - ny, & \rho_{22} &= b'w - Ny. \end{aligned}$$

It may readily be found that

$$\begin{aligned} \rho_{11v} &= (B_v + Bm)z - n_v x, & \rho_{12u} &= (b_u + bM)w - N_u x, \\ \rho_{21v} &= (B'_v + B'm)z - n_v y, & \rho_{22u} &= (b'_u + b'M)w - N_u y. \end{aligned}$$

Hence the developables of Γ_{ij} correspond to the developables of Γ and $\bar{\Gamma}$.

Denote by ρ_{ij}^λ the focal points other than ξ (or η) on the lines of the congruences Γ_{ij}^λ . We find that the coordinates of ρ_{ij}^λ are defined by the formulas

$$(43) \quad \begin{aligned} \rho_{11}^\lambda &= (B + \lambda B')z - nx, \\ \rho_{12}^\lambda &= (b + \lambda b')w - Nx. \end{aligned}$$

Hence

$$(44) \quad \rho_{11}^0 = \rho_{11}, \quad \rho_{12}^0 = \rho_{12}, \quad \rho_{11}^\infty = \rho_{21}, \quad \rho_{12}^\infty = \rho_{22}.$$

It follows from (43) that *each of the focal points of a line of a congruence of a pencil moves along a line as that congruence generates the pencil.* In the pencil as defined by Fubini in the paper cited, one focal point is fixed, the other focal point moves along a line.

8. The Transversal Surfaces

From (30) we may show that the coordinates x, y of the current points of S_x, S_y satisfy the following differential equations

$$\begin{aligned}
 (45) \quad x_{uu} &= \left(\frac{b_u}{b} - M \right) x_u + \frac{bs}{B} x_v + b(px + qy), \\
 x_{uv} &= \left(\frac{b_v}{b} + m \right) x_u + \left(\frac{B_u}{B} + M \right) x_v, \\
 x_{vv} &= \frac{BS}{b} x_u + \left(\frac{B_v}{B} - m \right) x_v + B(Qx + Py); \\
 (46) \quad y_{uu} &= \left(\frac{b'_u}{b'} - M \right) y_u + \frac{b's}{B} y_v + b'(qy + px), \\
 y_{uv} &= \left(\frac{b'_v}{b'} + m \right) y_u + \left(\frac{B'_u}{B'} + M \right) y_v, \\
 y_{vv} &= \frac{B'S}{b'} y_u + \left(\frac{B'_v}{B'} - m \right) y_v + B'(Py + Qx).
 \end{aligned}$$

Hence N_x, N_y are conjugate nets in the relation of a transformation F .

It follows from (45) and (46) that N_x and N_y have equal point invariants if the respective functions vanish:

$$(47) \quad E_x = \frac{\partial^2}{\partial u \partial v} \log \frac{b}{B}, \quad E_y = \frac{\partial^2}{\partial u \partial v} \log \frac{b'}{B'}.$$

Again from (45) and (46) we note that the asymptotic curves on S_x, S_y are given by the respective equations

$$(48) \quad bqdu^2 + Bpdv^2 = 0, \quad b'pdu^2 + B'Qdv^2 = 0.$$

But from (31)

$$(49) \quad \frac{bq}{Bp} = \frac{b'p}{B'Q}.$$

Hence the asymptotic curves on S_x, S_y correspond.

If we denote the coordinates of the tangent planes to $S_x, S_y, S_z, S_{\bar{z}}$ by ξ, η, ω, ζ respectively, we may write

$$\begin{aligned}
 (50) \quad \xi &= (x, z, w), & \eta &= (y, w, z), & \zeta &= (x, y, w) \\
 \omega &= (y, x, z).
 \end{aligned}$$

We find that these functions satisfy the following system of differential equations

$$\begin{aligned}
 (51) \quad \xi_u &= q\zeta, & \eta_u &= p\zeta, \\
 \xi_v &= P\omega, & \eta_v &= Q\omega, \\
 \xi_u &= b'\xi + b\eta + M\zeta - N\omega, & \omega_u &= -s\zeta - M\omega, \\
 \xi_v &= -m\zeta - S\omega, & \omega_v &= B'\xi + B\eta - n\zeta + M\omega.
 \end{aligned}$$

The equations of Laplace which ξ , η satisfy are

$$(52) \quad \begin{aligned} \xi_{uv} &= \left(\frac{q_v}{q} - m \right) \xi_u + \left(\frac{P_u}{P} - M \right) \xi_v, \\ \eta_{uv} &= \left(\frac{p_v}{p} - m \right) \eta_u + \left(\frac{Q_u}{Q} - M \right) \eta_v. \end{aligned}$$

It follows therefore that the nets N_x , N_y have equal tangential invariants if the respective invariants

$$(53) \quad E_\xi = \frac{\partial^2}{\partial u \partial v} \log \frac{q}{P}, \quad E_\eta = \frac{\partial^2}{\partial u \partial v} \log \frac{p}{Q}$$

vanish. From (47), (49), and (53) we see that

$$E_x + E_\xi = E_y + E_\eta.$$

As is seen from (33) the nets N_x , N_y are in the relation of a transformation K of Koenigs if and only if

$$(54) \quad bB' + b'B = 0.$$

Moreover from (31) and (54) one may show that

$$(55) \quad pP + qQ = 0.$$

But the condition (55) implies that the tangent plans to S_x and S_y at x, y separate the focal planes of the line g of Γ harmonically.

One may show from (32) that the condition (54) implies⁷ that $E_x = E_y = 0$, and that (55) implies that $E_\xi = E_\eta = 0$. It follows therefore that if the two nets N_x , N_y are in the relation of a transformation K they are also in the relation of a transformation⁷ Ω .

MICHIGAN STATE COLLEGE,
EAST LANSING, MICH.

⁷ L. P. Eisenhart, *Transformations of Surfaces*, Princeton, 1923, p. 134.

ON THE HOMOTOPY GROUPS OF SPHERES AND ROTATION GROUPS¹

BY GEORGE W. WHITEHEAD

(Received February 11, 1942)

1. Introduction

One of the outstanding problems in modern topology is that of classifying the mappings of an m -dimensional sphere S^m into a topological space X . In terms of the Hurewicz theory of homotopy groups² this problem may be phrased as follows: to determine the structure of the m^{th} homotopy group $\pi_m(X)$. Of particular interest is the case where X itself is an n -sphere S^n . In this case the results of Hopf,³ Freudenthal,⁴ and Pontrjagin⁵ have led to the solution of the problem for $m \leq n + 2$. For $m > n + 2$ almost nothing is known concerning the structure of $\pi_m(S^n)$.

That this problem is closely related to the study of homotopy properties of the rotation group R_n of the n -sphere has been shown by Pontrjagin,⁶ who has used the one- and two-dimensional homotopy groups of R_n to compute the groups $\pi_{n+i}(S_n)$ ($i = 1, 2$).

In the present paper we introduce an operation which associates with each mapping $f(S^m \times S^n) \subset S^n$ a mapping $\phi(S^{m+n+1}) \subset S^{n+1}$. This is a generalization of the procedure of Hopf⁶ for the case $m = n$. This operation is shown to induce a homomorphism of $\pi_m(R_n)$ into $\pi_{m+n+1}(S^{n+1})$, which for $m = 1, 2$ turns out to be an isomorphism. The connection of this homomorphism with one introduced by Freudenthal⁴ is studied.

In a recent paper Freudenthal⁷ has announced without proof a very general theorem on extension of mappings, and used this theorem to construct maps of S^{2n-1} on S^n of Hopf invariant 1⁶ for all even n . We shall use the above results to construct a counter-example to Freudenthal's theorem. It is further shown that Freudenthal's construction definitely fails if $n > 2$ and $n \equiv 2 \pmod{4}$.

2. Preliminary concepts

In Euclidean $(r + 1)$ -space \mathbb{E}^{r+1} let S^r denote the unit sphere, i.e., the set of points $x = (x_1, \dots, x_{r+1}) \in \mathbb{E}^{r+1}$ with

$$(1) \quad |x|^2 = \sum_{i=1}^{r+1} x_i^2 = 1.$$

¹ Presented to the American Mathematical Society, December 30, 1941.

² W. Hurewicz, Proc. Akad. Amsterdam 38 (1935), pp. 112-119.

³ H. Hopf, Math. Ann. 104 (1931), pp. 637-665. We shall refer to this paper as H I.

⁴ H. Freudenthal, Comp. Math. 5 (1937), pp. 299-314. We shall refer to this paper as F I.

⁵ L. Pontrjagin, C. R. Acad. Sci. URSS 19 (1938), pp. 147-149, 361-363.

⁶ H. Hopf, Fund. Math. 25 (1935), pp. 427-440. We shall refer to this paper as H II.

⁷ H. Freudenthal, Proc. Akad. Amsterdam 42 (1939), pp. 139-140. We shall refer to this paper as F II.

Let E_i^r ($i = 1, 2$) be the hemispheres defined by the conditions $x_{r+1} \geq 0$, $x_{r+1} \leq 0$, respectively. E^{r+1} denotes the closed $(r+1)$ -cell $|x| \leq 1$ bounded by S^r . We shall refer to the points $x^1 = (0, 0, \dots, 1)$ and $x^2 = (0, 0, \dots, -1)$ as the *north* and *south poles*, respectively.

Let Y be a metric space with distance function $\rho(y_1, y_2)$, y^0 a fixed point of Y . By Y^{S^r} we shall mean the space of all mappings⁸ $f(S^r) \subset Y$ metrized by

$$(2) \quad \rho(f, g) = \text{L.U.B.}_{x \in S^r} \rho[f(x), g(x)] \quad (f, g \in Y^{S^r}).$$

Let x^0 be the point of S^r with co-ordinates $(1, 0, \dots, 0)$. Then $Y^{S^r}(x^0, y^0)$ denotes the subspace of Y^{S^r} consisting of those mappings $f(S^r) \subset Y$ such that $f(x^0) = y^0$. Two mappings $f, g \in Y^{S^r}(x^0, y^0)$ are said to be homotopic if they can be joined by an arc in $Y^{S^r}(x^0, y^0)$. The relation of homotopy is reflexive, symmetric, and transitive and divides the space $Y^{S^r}(x^0, y^0)$ into equivalence classes, called *homotopy classes*. The set of all these homotopy classes we denote by $\pi_r(Y)$. We shall denote the homotopy class of any $f \in Y^{S^r}(x^0, y^0)$ by \mathbf{f} .

We define an operation of addition between homotopy classes as follows: let f_i ($i = 1, 2$) $\in Y^{S^r}(x^0, y^0)$. Let ϕ_i ($i = 1, 2$) be a mapping of E_i^r on S^r such that (1) $\phi_i(S^{r-1}) = x^0$; (2) $\phi_i(E_i^r - S^{r-1}) \subset S^r$ is a topological map of degree 1. Then we define a mapping $f(S^r) \subset Y$ as follows:

$$(3) \quad f(x) = \begin{cases} f_1[\phi_1(x)] & (x \in E_1^r), \\ f_2[\phi_2(x)] & (x \in E_2^r). \end{cases}$$

It is easily verified that the homotopy class of f depends only on the homotopy classes of f_1 and f_2 . Let

$$(4) \quad \mathbf{f} = \mathbf{f}_1 + \mathbf{f}_2.$$

Hurewicz² has proved that under the operation of addition so defined the set $\pi_r(Y)$ becomes a group, called the r^{th} *homotopy group* of Y . This group is abelian if $r > 1$; in all the cases we consider here it is also abelian if $r = 1$.

3. The homomorphism H

Let Euclidean $(m+n+2)$ -space be represented as the product space $\mathfrak{E}^{m+1} \times \mathfrak{E}^{n+1}$, points $x \in \mathfrak{E}^{m+n+2}$ being represented by co-ordinates (p, q) ($p \in \mathfrak{E}^{m+1}$, $q \in \mathfrak{E}^{n+1}$). Then S^{m+n+1} is defined by

$$(5) \quad |p|^2 + |q|^2 = 1.$$

Let H_1 and H_2 be the subsets of S^{m+n+1} defined by

$$(6_1) \quad |p| \leq |q|,$$

$$(6_2) \quad |p| \geq |q|,$$

⁸ All mappings are supposed continuous.

respectively. Let

$$(7_1) \quad \psi_1(p, q) = (p/|q|, q/|q|) \quad ((p, q) \in H_1),$$

$$(7_2) \quad \psi_2(p, q) = (p/|p|, q/|p|) \quad ((p, q) \in H_2).$$

Evidently $\psi_1|_{H_1H_2} = \psi_2|_{H_1H_2}$ and maps H_1H_2 into $S^m \times S^n$. Denote this mapping by ψ . Then

LEMMA 1. *The mappings ψ_1 , ψ_2 , and ψ defined above are homeomorphic mappings of H_1 on $E^{m+1} \times S^n$, H_2 on $S^m \times E^{n+1}$, and H_1H_2 on $S^m \times S^n$ respectively.*

Let f be a mapping of $S^m \times S^n$ into S^n . We associate with f the mapping $H(f) = \phi(S^{m+n+1}) \subset S^{n+1}$ as follows: ϕ maps the great circle joining the point $(0, q)$ to the point (p, q) on the great circle joining the north pole z^1 of S^{n+1} to the point $f[\psi^{-1}(p, q)]$, and maps the great circle joining $(p, 0)$ to (p, q) on the great circle joining z^2 to $f[\psi^{-1}(p, q)]$. Evidently $\phi(H_1) \subset E_1^{n+1}$, $\phi(H_2) \subset E_2^{n+1}$, while $\phi = f\psi^{-1}$ on H_1H_2 . The functions defining the mapping ϕ are given by

$$(8) \quad \begin{aligned} \phi_i(p, q) &= 2|p| \cdot |q| \cdot f_i(p/|p|, q/|q|) & (|p| \cdot |q| \neq 0), \\ \phi_i(0, q) &= \phi_i(p, 0) = 0 & (i = 1, \dots, n+1); \\ \phi_{n+2}(p, q) &= |q|^2 - |p|^2. \end{aligned}$$

We use this operation to construct a mapping $H = H_{m,n}$ of $\pi_m(R_n)$ into $\pi_{m+n+1}(S^{n+1})$ as follows: let $e \in R_n$ denote the identity mapping of S^n on itself, and let $f \in R_n^{\text{sm}}(p^0, e)$. If $p \in S^m$, $q \in S^n$, let $f^*(p, q)$ denote the point of S^n into which q is carried by the rotation $f(p)$. Let $\phi = H(f^*)$. Then it is easy to verify that $\phi \in S^{n+1, sm+n+1}(x^0, z^2)$, where $x^0 = (p^0, 0)$ and z^2 is the south pole of S^{n+1} . Let $H(f) = \phi$. Evidently $f = g$ implies $H(f) = H(g)$, so that H is a well-defined mapping of $\pi_m(R_n)$ into $\pi_{m+n+1}(S^{n+1})$. We have further

THEOREM 1. *H is a homomorphic mapping of $\pi_m(R_n)$ into $\pi_{m+n+1}(S^{n+1})$.*

For let $f, g \in \pi_m(R_n)$, and let h be the constant mapping $h(p) = e$ ($p \in S^m$). Then $h = 0$. Hence $f + h = f$, $h + g = g$, so that $H(f + h) = H(f)$, $H(h + g) = H(g)$. It is therefore sufficient to prove that

$$(9) \quad H(f + h) + H(h + g) = H(f + g).$$

Let f', g' be mappings of S^m into R_n defined by

$$(10_1) \quad \begin{aligned} f'(p) &= f[\phi_1(p)] & (p \in E_1^m), \\ &= h[\phi_2(p)] & (p \in E_2^m); \end{aligned}$$

$$(10_2) \quad \begin{aligned} g'(p) &= h[\phi_1(p)] & (p \in E_1^m), \\ &= g[\phi_2(p)] & (p \in E_2^m). \end{aligned}$$

Then $f' = f + h$, $g' = h + g$. Let $F = H(f'^*)$, $G = H(g'^*)$.

* If $f(x) \subset Y$ and A is a closed subset of X , $f|A$ denotes the mapping of A into Y obtained by restricting the range of definition of f to the set A .

Let π_i denote the vertical projection of E_i^{m+n+1} on E^{m+n+1} ($i = 1, 2$). Then $\pi_i(x) = x$ for $x \in S^{m+n}$. Let $F_0 = F|E_1^{m+n+1}$, $H'' = F|E_2^{m+n+1}$, $H' = G|E_1^{m+n+1}$, $G_0 = G|E_2^{m+n+1}$. Then it is easily verified that $H'\pi_1^{-1} = H''\pi_2^{-1}$. Call this mapping H_0 . Evidently $F_0(x) = G_0(x) = H_0(x)$ ($x \in S^{m+n}$).

Let H_t ($0 \leq t \leq 1$) be a homotopy of H_0 to x^0 keeping x^0 fixed. Then¹⁰ there exist homotopies F_t, G_t ($0 \leq t \leq 1$) of F_0, G_0 respectively, such that $F_t(x) = G_t(x) = H_t(x)$ ($x \in S^{m+n}$). Let

$$(11_1) \quad \begin{aligned} F'_t(x) &= F_t(x) & (x \in E_1^{m+n+1}), \\ &= H_t[\pi_2(x)] & (x \in E_2^{m+n+1}); \end{aligned}$$

$$(11_2) \quad \begin{aligned} G'(x) &= H_t[\pi_1(x)] & (x \in E_1^{m+n+1}), \\ &= G_t(x) & (x \in E_2^{m+n+1}). \end{aligned}$$

Evidently $F'_1 = F, G'_1 = G$.

Let

$$(12) \quad \begin{aligned} H'_t(x) &= F_t(x) & (x \in E_1^{m+n+1}), \\ &= G_t(x) & (x \in E_2^{m+n+1}). \end{aligned}$$

Then $H'_0 = H(f + g)$, while $H'_1 = F'_1 + G'_1 = F + G = H(f + h) + H(f + g)$.¹¹ But $H'_0 = H'_1$, which proves the theorem.

4. Relations between the homomorphisms F, G , and H

Let S^{m+n} be the equator of S^{m+n+1} , S^n the equator of S^{n+1} , and let f be a mapping of S^{m+n} into S^n . We associate with the mapping f a mapping $F(f) = \psi(S^{m+n+1}) \subset S^{n+1}$ as follows: ψ maps the great circle joining the north pole x^1 of S^{m+n+1} to the point $x \in S^{m+n}$ on the great circle joining z^1 to $f(x)$, and maps the great circle joining x^2 to x on the great circle joining z^2 to $f(x)$. Evidently $\psi(E_1^{m+n+1}) \subset E_1^{n+1}$, $\psi(E_2^{m+n+1}) \subset E_2^{n+1}$, while $\psi = f$ on S^{m+n} . If $f \in S^{n, S^{m+n+1}}(x^0, y^0)$, then $F(f) \in S^{n+1, S^{m+n+1}}(x^0, y^0)$; moreover, f homotopic to g implies $F(f)$ homotopic to $F(g)$. Thus F induces a mapping F of $\pi_{m+n}(S^n)$ into $\pi_{m+n+1}(S^{n+1})$, which was shown by Freudenthal⁴ to be a homomorphism.

Let R_{n-1} be the closed subgroup of R_n consisting of those rotations which leave the north pole fixed. Evidently R_{n-1} is isomorphic with the group of rotations of S^{n-1} . Since $R_{n-1} \subset R_n$, there is a natural homomorphism G of $\pi_m(R_{n-1})$ into $\pi_m(R_n)$.

THEOREM 2. *The homomorphisms F, G , and H are related by*

$$(13) \quad FH_{m,n-1} = H_{m,n}G.$$

¹⁰ K. Borsuk, *Fund. Math.* 28 (1937), p. 101.

¹¹ This follows from the definition of addition in $\pi_{m+n+1}(S^{n+1})$ given by S. Eilenberg (*Ann. of Math.* 41 (1940), p. 235), which is easily shown to be equivalent to the one given here.

For let $f \in \pi_m(R_{m-1})$, $g = F[H_{m,n-1}(f)]$, $g' = H_{m,n}[G(f)]$. It is then easily verified that $g = g'$ on S^{m+n} . Moreover $g'(E_1^{m+n+1}) \subset E_1^{n+1}$, $g'(E_2^{m+n+1}) \subset E_2^{n+1}$. Hence for no x is $g'(x) = -g(x)$. It follows that g and g' are homotopic, so that $g = g'$.

Let ϕ be a mapping of S^{n-1} into R_{n-1} defined as follows: if $x \in S^{n-1}$, x' is the point in the great circle joining x^1 to x whose angular distance from x^1 is twice that from x^1 to x . Then $\phi(x)$ is that rotation which carries x^1 into x' and leaves each point in the $(n-2)$ -sphere orthogonal to x^1 and x fixed. Let $h = H_{n-1,n-1}(\phi)$. Then it can easily be shown¹² that if n is even h has Hopf invariant 2. We have further:

THEOREM 3. *The kernel of the homomorphism $F[\pi_{2n-1}(S^n)] \subset \pi_{2n}(S^{n+1})$ (n even) is the subgroup of $\pi_{2n-1}(S^n)$ generated by h .*

The author has recently shown¹³ that $G(\phi) = 0$; in fact, the kernel of the homomorphism G is the subgroup of $\pi_{n-1}(R_{n-1})$ generated by ϕ . It follows from Theorem 2 that $F[H_{n-1,n-1}(\phi)] = F(h) = 0$. Let $g \in \pi_{2n-1}(S^n)$, and suppose that $F(g) = 0$. Then the Hopf invariant of g is even,¹⁴ say $2k$. Let $f = kh$. Then $F(f - g) = 0$, and $f - g$ has Hopf invariant zero. Hence¹⁵ $f - g = 0$, i.e., $g = f = kh$.

THEOREM 4. *$H_{m,n}$ maps $\pi_m(R_n)$ isomorphically for $m = 1, 2$. $H_{m,n}$ maps $\pi_m(R_n)$ on $\pi_{m+n+1}(S^{n+1})$ for $m = 1$ and for $m = 2, n > 1$.*

Let $h(S^1) \subset R_1$ be defined by

$$h(x) = \begin{vmatrix} x_1 & -x_2 \\ x_2 & x_1 \end{vmatrix}.$$

Then h maps S^1 homeomorphically on R_1 , and h is a generator of the free cyclic group $\pi_1(R_1)$. But $H_{1,1}(h)$ maps S^3 on S^2 with Hopf invariant 1¹⁶ and generates the group $\pi_3(S^2)$. It follows from Theorems 2 and 3 that $H_{1,n}$ maps $\pi_1(R_n)$ isomorphically on $\pi_{n+2}(S^{n+1})$ for $n > 1$.

Since $\pi_2(R_n) = 0$, it follows that $H_{2,n}$ is an isomorphism. But $\pi_{n+3}(S^{n+1}) = 0$ for $n > 1$ ⁶, and hence $H_{2,n}$ maps $\pi_2(R_n)$ on $\pi_{n+3}(S^{n+1})$. This completes the proof of the theorem.

5. Freudenthal's theorem

Freudenthal has recently announced⁷ without proof a very general theorem on extension of mappings, and used this theorem to construct maps of S^{2n-1} on S^n with Hopf invariant 1 for all even n .¹⁷ In this section the foregoing results are used to construct a counter-example to Freudenthal's theorem, and to show that the above-mentioned construction fails if $n > 2$ and $n \equiv 2 \pmod{4}$.

¹² Cf. H II, p. 431.

¹³ Ann. of Math. 43 (1942), Theorem 5.

¹⁴ F I, Satz III.

¹⁵ F I, Satz II, 2.

¹⁶ H I, p. 654.

¹⁷ F II, p. 140.

Let points z of Euclidean $2n$ -space be represented by complex co-ordinates (z_1, \dots, z_n) . Then S^{2n-1} is represented by the equation $\sum_{i=1}^n z_i \bar{z}_i = 1$.

Let P_{n-1} denote complex projective $(n-1)$ -space. Then there is a natural mapping $\phi(S^{2n-1}) \subset P_{n-1}$ defined by mapping each point $z \in S^{2n-1}$ into the point of P_{n-1} with the same coordinates. This is evidently a fibre map in the sense of Hurewicz and Steenrod,¹⁸ the fibres being great circles. This mapping $\phi(S^{2n-1}) \subset P_{n-1}$ can be extended to a mapping $\psi(E^{2n}) \subset P_n$, where $\psi(z_1, \dots, z_n) = (z_1, \dots, z_n, (1 - \sum z_i \bar{z}_i)^{1/2})$. It is easily verified that ψ is a homeomorphism on $E^{2n} - S^{2n-1}$ and $\psi = \phi$ on S^{2n-1} .

Let X be a topological space, f a mapping of P_{n-1} into X . Then

THEOREM 5. *The mapping $f(P_{n-1}) \subset X$ can be extended to a mapping $f^*(P_n) \subset X$ if and only if the mapping $\phi(S^{2n-1}) \subset X$ is inessential.*

For if ϕ is inessential, there is a mapping $F(E^{2n}) \subset X$ such that $F = \phi$ on S^{2n-1} . Let $f^* = F\psi^{-1}$. Then f^* is the required extension. Conversely, if f^* is an extension of f , let $F = f^*\psi$. Then F maps E^{2n} into X and $F = \phi$ on S^{2n-1} . Hence ϕ is inessential.

Let $g(S^1) \subset R_{2n-1}$ be defined by

$$g(x) = \begin{pmatrix} x_1 & x_2 & 0 & 0 & \cdots & 0 & 0 \\ -x_2 & x_1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & x_1 & x_2 & \cdots & 0 & 0 \\ 0 & 0 & -x_2 & x_1 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \cdots & x_1 & x_2 \\ 0 & 0 & 0 & 0 & \cdots & -x_2 & x_1 \end{pmatrix}.$$

Then g is essential or inessential according as n is odd or even. For if $n = 1$, g is a generator of $\pi_1(R_1)$, so that g is essential. If $n = 2$, we have $g(S^1) \subset Q^3$, where Q^3 is the quaternion subgroup of R_3 . But $\pi_1(Q^3) = \pi_1(S^3) = 0$. Hence $g = 0$ in $Q^3 \subset R_3$, and g is inessential. The proof is completed by induction.

Let $h = H(g)$. Then it follows from Theorem 4 that $h(S^{2n+1}) \subset S^{2n}$ is essential if n is odd and inessential if n is even. Moreover, it can be directly verified that there is a mapping $h'(P_n) \subset S^{2n}$ such that $h = h'\phi$, and that h' has degree 1. An application of Theorem 5 gives

THEOREM 6. *If n is even, the mapping $h'(P_n) \subset S^{2n}$ can be extended over P_{n+1} . If n is odd, it cannot be so extended.*

The theorem of Freudenthal's referred to above can be phrased as follows:¹⁹ Let K be a complex, f a normal mapping²⁰ of K^q into S^q . Suppose that f can be extended over K^{q+1} . Then f can be extended over K^{2q-1} .

Let K be a triangulation of P_{n+1} , so that P_n becomes a closed subcomplex L of K . Then $L \subset K^{2n}$. Let h' be the mapping of L into S^{2n} of degree one

¹⁸ W. Hurewicz and N. E. Steenrod, Proc. Nat. Acad. 27 (1941), pp. 60-64.

¹⁹ F II, p. 140.

²⁰ I.e., $f(K^{q-1}) = x^0$.

described above. Then²¹ h' can be deformed into a normal map h'' ; moreover, h'' can be extended over K if and only if the same is true of h' . Let $H^r(K - L)$ denote the r^{th} cohomology group of $K - L$ with integral coefficients. Then $H^r(K - L) = 0$ for $r < 2n + 2$, while $H^{2n+2}(K - L)$ is a free cyclic group. In particular, $H^{2n+1}(K - L) = 0$. It follows from a theorem of Whitney²² that h'' can be extended over K^{2n+1} . But h'' cannot be extended over K^{2n+2} for n odd.

Freudenthal's construction of maps of S^{4n-1} on S^{2n} is based on an application of his theorem to the case $K = P_{2n}$, $f(K^{2n}) \subset S^{2n}$, where $f(P_n) \subset S^{2n}$ is of degree one. The argument above shows that this construction breaks down if n is odd and > 1 ; for f cannot even be extended over the subspace P_{n+1} of P_{2n} .

PURDUE UNIVERSITY

²¹ H. Whitney, Duke Journal 3 (1937), p. 53.

²² Loc. cit., Theorem 2.

LINEAR p -ADIC GROUPS AND THEIR LIE ALGEBRAS

By ROBERT HOOKE

(Received December 18, 1941)

1. Introduction

The theory of real or complex Lie groups necessarily treats only those topological groups which are locally connected. It is the object of this paper to develop methods of Lie theory for the opposite case of totally disconnected groups.

We shall consider those groups into which can be introduced local analytic coordinates from a p -adic field K , calling these p -adic Lie groups over K . The concept of the associated Lie algebra over K will be used at once to obtain the usual properties of Lie groups. Bearing in mind the results of Ado¹ on imbedding any Lie algebra of characteristic zero in a Lie algebra of matrices, we shall restrict ourselves to the study of Lie algebras of matrices over K and the p -adic Lie groups contained in the full linear group over K .

Although the coordinates in these groups may be defined only in a certain neighborhood of the identity, all the groups which will occur will be entire groups. It is necessary, however, to identify two subgroups whose intersection is open in each, as is done in the theory of real local Lie groups. It will then be shown in sections 6 and 7 that the usual one-to-one correspondence exists between the subgroups of a Lie group and the subalgebras of its Lie algebra.

The last sections are devoted to certain special groups and their Lie algebras, and in particular to the determination of groups whose Lie algebras are the various "non-exceptional" normal simple Lie algebras over K , which have been classified by Jacobson.

The author wishes to express here his appreciation of the assistance and encouragement of Professor C. Chevalley in the preparation of this paper.

2. Notation and Preliminary Theorems

We shall first list, without proof, a few necessary theorems from p -adic analysis. Unless otherwise noted, these theorems may be found in the papers of Chabauty, [4], and Chevalley, [5].

Let R be the field of rational numbers and p be a fixed prime. There is determined by p a valuation v in R which is defined by

$$v(p) = \rho \quad 0 < \rho < 1, \rho \text{ a real number.}$$

(For this notation and elementary results, the reader is referred to Albert, [2].) R_p will denote the complete p -adic number field determined from R by the valuation v . This valuation has the properties:

- (a) $v(xy) = v(x)v(y),$
- (b) $v(x + y) \leq \max v(x), v(y).$

¹ See Ado, [1]. The numbers in square brackets will refer to papers in the bibliography.

If x is an element of R_p , then $v(x)$ is defined and is equal to p^m where m is some rational integer. If m is not negative, that is, if $v(x) \leq 1$, then x is called an integer of R_p . The integers of R_p form a ring E_{R_p} all of whose ideals are powers of the prime ideal (p) .

Now let K be any finite algebraic extension of R_p . The integers of K are defined as those elements whose irreducible equations over R_p have coefficients which are all integers of R_p . The valuation v has a unique extension to a valuation of K , and the integers of K are those elements k such that $v(k) \leq 1$.

If we put $d(x, y) = v(x - y)$ in K , then K becomes a metric space which is complete, locally compact, and totally disconnected. The symbol K^n will denote the direct product space of K by itself to n factors and will be called p -adic n -space over K .

A series $\sum a_n$ with terms in K converges if and only if $\lim_{n \rightarrow \infty} v(a_n) = 0$. An analytic function defined on K^n is by definition the sum in its region of convergence of a power series of the type

$$\sum a_{h_1 h_2 \dots h_n} x_1^{h_1} x_2^{h_2} \dots x_n^{h_n}$$

which converges for all points (x_1, x_2, \dots, x_n) in some neighborhood of the origin in K^n . Such a power series can be differentiated term by term to give a new series convergent in the same region. It has the properties of the absolutely convergent series of complex analysis.

We conclude this summary with a list of theorems from p -adic analysis. These will be referred to throughout the paper by the numbers given them.

(1) Every integer of R_p is a limit of rational integers.

(2) Let $f(x_1, x_2, \dots, x_n)$ be an analytic function defined in a neighborhood D of the origin in K^n . Let $f_i(y_1, y_2, \dots, y_m)$, $(i = 1, 2, \dots, n)$ be n analytic functions defined in a neighborhood D' of the origin in K^m such that $f_i(0, 0, \dots, 0) = 0$, $(i = 1, 2, \dots, n)$. Then if in f we substitute for each x_i the series f_i , we get a series $f'(y_1, y_2, \dots, y_m)$ which converges for all points y in D' which are such that the point with coordinates $f_i(y_1, y_2, \dots, y_m)$ is in D .

(3) Let f_1, f_2, \dots, f_h be h functions of $h + m$ variables $u_1, u_2, \dots, u_h; x_1, x_2, \dots, x_m$ such that

$$f_i(0, 0, \dots, 0) = 0, \quad i = 1, 2, \dots, h,$$

and such that

$$\frac{D(f_1, f_2, \dots, f_h)}{D(u_1, u_2, \dots, u_h)} \neq 0$$

when the u_i and the x_i are all zero. Then the equations

$$f_i(u_1, u_2, \dots, u_h; x_1, x_2, \dots, x_m) = 0, \quad i = 1, 2, \dots, h,$$

define the u_i as analytic functions of the x_i ,

$$u_i = F_i(x_1, x_2, \dots, x_m), \quad i = 1, 2, \dots, h,$$

where

$$F_i(0, 0, \dots, 0) = 0, \quad i = 1, 2, \dots, h.$$

(4) (Cf. Lutz, [14].) A system of differential equations

$$dy_i/dt = f_i(t, y_1, y_2, \dots, y_n), \quad i = 1, 2, \dots, n,$$

where the f_i are analytic functions, has, in some neighborhood of $t = y_i = 0$, one and only one solution in the form

$$y_i = g_i(t), \quad i = 1, 2, \dots, n,$$

where the g_i are analytic functions with the initial conditions

$$g_i(0) = 0, \quad i = 1, 2, \dots, n.$$

(5) If x is an element of K and $v(x) < \rho^{1/(p-1)}$, the series

$$\exp x = 1 + x + x^2/2! + \dots + x^n/n! + \dots$$

converges, and $v(\exp x - 1) = v(x)$. We have, as in ordinary analysis, $\exp(x + y) = (\exp x)(\exp y)$, when all of these exist.

(6) If x is an element of K , the series

$$\log x = (x - 1) - (x - 1)^2/2 + \dots + (-1)^{n-1}(x - 1)^n/n + \dots$$

converges when $v(x - 1) < 1$.

(7) If $v(x) < \rho^{1/(p-1)}$, $\log(\exp x)$ exists and is equal to x . If $v(x - 1) < \rho^{1/(p-1)}$, $\exp(\log x)$ exists and is equal to x . To prove the second statement, we need only show that $v(\log x) < \rho^{1/(p-1)}$. We have

$$v(\log x) \leq \max(v_n), \quad \text{where } v_n = v[(x - 1)^n/n].$$

If $v(n) = \rho^\alpha$, then $n \geq p^\alpha$, and we have

$$\alpha \leq p^{\alpha-1} + p^{\alpha-2} + \dots + p + 1 = (p^\alpha - 1)/(p - 1) \leq (n - 1)/(p - 1).$$

Hence $v_n < (\rho^{1/(p-1)})^n / (\rho^{(n-1)/(p-1)}) \leq \rho^{1/(p-1)}$. Q.E.D.

(8) If an analytic function $f(t)$ is equal to zero for a sequence of values of t approaching zero and $f(0) = 0$, then $f(t)$ is identically zero. (The proof is as in ordinary analysis.)

3. Matrices in a p -adic Field

Given a field K , we shall denote by K_n the full matrix algebra of n -rowed square matrices over K , and by $K_{n,1}$ the Lie algebra obtained from K_n by defining the commutator by

$$[x, y] = xy - yx.$$

This will be called a pure commutator of degree 2 in x and y . If c^m is a pure commutator of degree m in x and y , then by definition, $[c^m, x]$ and $[c^m, y]$ are pure commutators of degree $m + 1$ in x and y . The full linear group of order n over K we shall denote simply by G_K , since n will be fixed throughout.

A sequence of matrices is said to converge, if for each fixed pair i, j , the elements in the i^{th} row and j^{th} column of these matrices form a convergent sequence. The limit matrix is the matrix whose elements are the limits of these sequences.

Given a matrix $A = ||a_{ij}||$ in K_n , there can be defined a weak type of valuation V in K_n by putting $V(A) = \max v(a_{ij})$. This valuation satisfies the conditions

$$(a') \quad V(A + B) \leq \max V(A), V(B),$$

$$(b') \quad V(tA) = v(t)V(A), \quad t \text{ in } K,$$

$$(c') \quad V(AB) \leq V(A)V(B).$$

Now a sequence of matrices $A_m = ||a_{ij}^{(m)}||$ with elements in a p -adic field K converges if and only if

$$\lim_{m \rightarrow \infty} v(a_{ij}^{(m)} - a_{ij}^{(m+1)}) = 0 \text{ for all } i, j.$$

We have, however,

$$v(a_{ij}^{(m)} - a_{ij}^{(m+1)}) \leq V(A_m - A_{m+1})$$

for each i and j , and there exists for each m a pair of integers i and j such that the equality sign holds. We thus get the same condition for the convergence of a sequence of matrices as for a sequence of elements in K , using now the valuation V . It can be seen, therefore, that in dealing with sequences and series of matrices, we get automatically the same theorems for commutative matrices with the valuation V that are proved for elements in K with the valuation v , provided these theorems are proved using only the fact that the valuation v satisfies the conditions (a'), (b'), (c').

In particular, we have the following theorems.

THEOREM 1. *If X is in K_n and $V(X) < \rho^{1/(p-1)}$, the series*

$$\exp tX = I + tX + \cdots + t^n X^n/n! + \cdots \quad (I \text{ the identity matrix})$$

converges for $v(t) \leq 1$ and we have $V(\exp tX - 1) = V(tX)$. Also for any matrix Y there exists a real number $r > 0$ such that $\exp tY$ converges for $v(t) \leq r$.

THEOREM 2. *If $V(A - I) < 1$, the series*

$$\log A = (A - I) - (A - I)^2/2 + \cdots + (-1)^{n-1}(A - I)^n/n + \cdots$$

converges. If $V(X) < \rho^{1/(p-1)}$, then $\log(\exp X)$ is defined and equal to X . Also if $V(A - I) < \rho^{1/(p-1)}$, then $\exp(\log A)$ is defined and equal to A .

If X and Y are commutative matrices, it can be shown in the usual way that $\exp(X + Y) = (\exp X)(\exp Y)$ when these exist. We shall need the following generalization of this fact for non-commutative matrices:

THEOREM 3. *Let X and Y be matrices in K_n such that*

$$V(X), V(Y) < \rho^{1/(p-1)}.$$

i) *There is a matrix Z defined by $\exp Z = (\exp X)(\exp Y)$.*

ii) $Z = X + Y + \lim_{m \rightarrow \infty} f_m$, where the f_m are linear combinations of higher commutators of X and Y with rational coefficients.

PROOF. i) By Theorem 1, $V((\exp X) - I) = V(X)$ and $V((\exp Y) - I) = V(Y)$. Hence

$$\begin{aligned} V[(\exp X)(\exp Y) - I] &= V[(\exp X - I)(\exp Y - I) \\ &\quad + (\exp X - I) + (\exp Y - I)] \\ &\leq \max V(X), V(Y). \end{aligned}$$

By Theorem 2, therefore, $Z = \log [(\exp X)(\exp Y)]$ exists and $\exp Z = (\exp X)(\exp Y)$.

ii) If X and Y are real matrices, it has been shown by Hausdorff, [7], that

$$Z = X + Y + \sum_{r=2}^{\infty} \sum_{s=1}^{\alpha_r} a^r f^{rs}(X, Y),$$

where f^{rs} is a pure commutator of degree r in X and Y and each a^r is a rational number. Each α_r is finite.

It follows that

$$Z = X + Y + \lim_{m \rightarrow \infty} \sum_{r=2}^m \sum_{s=1}^{\alpha_r} a^r f^{rs}(X, Y).$$

Each f^{rs} can be expressed as a polynomial which is homogeneous of degree r in X and Y . Let us now put

$$F^r(X, Y) = \sum_{s=1}^{\alpha_r} a^r f^{rs}(X, Y).$$

Then $Z = X + Y + \lim_{m \rightarrow \infty} \sum_{r=2}^m F^r(X, Y)$, where each F^r is a homogeneous polynomial of degree r in X and Y .

Expressing this series as n^2 series in the elements of the matrices therein, we have

$$\begin{aligned} Z_{ij} &= X_{ij} + Y_{ij} + \lim_{m \rightarrow \infty} \sum_{r=2}^m F_{ij}^r(X_{11}, \dots, X_{nn}; Y_{11}, \dots, Y_{nn}), \\ &\quad i, j = 1, 2, \dots, n. \end{aligned}$$

It is also true, however, that for sufficiently small values of the elements, the series

$$Z_{ij} = \{\log [(\exp X)(\exp Y)]\}_{ij}, \quad i, j = 1, 2, \dots, n,$$

converge. These may be written

$$\begin{aligned} Z_{ij} &= X_{ij} + Y_{ij} + \lim_{m \rightarrow \infty} \sum_{r=2}^m P_{ij}^r(X_{11}, \dots, X_{nn}; Y_{11}, \dots, Y_{nn}) \\ &\quad i, j = 1, 2, \dots, n \end{aligned}$$

where P^r is the sum of terms of degree r in the expansion. We now have the equivalent of two power series expansions for Z_{ij} in terms of the n^2 real variables X_{ij} , Y_{ij} . The two series must therefore be identical term by term.

If now X and Y are p -adic matrices, it follows from i) that

$$Z_{ij} = X_{ij} + Y_{ij} + \lim_{m \rightarrow \infty} \sum_{r=2}^m P_{ij}^r, \quad i, j = 1, 2, \dots, n,$$

and so

$$Z_{ij} = X_{ij} + Y_{ij} + \lim_{m \rightarrow \infty} \sum_{r=2}^m F_{ij}^r, \quad i, j = 1, 2, \dots, n.$$

Returning to the matrix expressions, we have

$$\begin{aligned} Z &= X + Y + \lim_{m \rightarrow \infty} \sum_{r=2}^m F^r(X, Y) \\ &= X + Y + \lim_{m \rightarrow \infty} \sum_{r=2}^m \sum_{s=1}^{\alpha_r} a'^s f^s(X, Y), \end{aligned}$$

since for any finite value of m these are identical.

Q.E.D.

4. p -adic Lie Groups

We make the following definition as in the theory of real and complex Lie groups:

DEFINITION. A p -adic Lie group G is a topological group equipped with a homeomorphic mapping of a neighborhood N of its identity element onto a neighborhood of the origin of K^m which satisfies the condition: If X, Y, Z are in N , $XY = Z$, and X is mapped on (x_1, x_2, \dots, x_m) in K^m , etc., then

$$z_i = f_i(x_1, x_2, \dots, x_m; y_1, y_2, \dots, y_m), \quad i = 1, 2, \dots, m,$$

where the f_i are analytic functions. G is then said to have local analytic coordinates.

The group G_K can be shown to be a p -adic Lie group. We introduce a metric topology by defining the distance between two elements A and B as $V(A - B)$. Then if we put each matrix A in the form $A = I + ||a_{ij}||$, we can define the $n^2 p$ -adic numbers a_{ij} as the coordinates of A . There clearly exists a real number $q > 0$ such that for any set of a_{ij} in K satisfying the inequality $v(a_{ij}) \leq q$, the matrix whose coordinates are these numbers is non-singular and in G_K . It follows that there exists a neighborhood of I in G_K homeomorphic to a neighborhood of the origin in K^n . Since multiplication is a polynomial operation in this group, these coordinates are analytic. Since K is locally compact, so is G_K .

A complete system of neighborhoods of I is furnished by the set of spheres S_r consisting of those matrices whose coordinates satisfy for some fixed real number r the inequality

$$v(a_{ij}) \leq r.$$

Since v is a discrete valuation, these spheres are both open and closed.

For $r < 1$, S_r is always a subgroup. From the nature of the valuation, the product of two elements in S_r is again in S_r . The existence of inverses follows from the fact that the determinant of any matrix in S_r has value 1. These spheres form an infinite descending chain of open and closed subgroups whose intersection is I , so G_K is totally disconnected. It is the existence of these open subgroups which creates most of the difference between the theory of p -adic Lie groups and that of real Lie groups.

5. The Lie Algebras of G_K and its Subgroups

The Lie algebra (infinitesimal group) of a p -adic Lie group may be defined exactly as in the case of real Lie groups (cf. Pontrjagin, [15]). Many relations here may be proved exactly as in the ordinary case, and so their proofs will be omitted. We shall depart from the usual procedure, however, by using the Lie algebra to obtain conditions for the subgroups of a Lie group to be themselves Lie groups.

An *analytic curve* in G_K is an analytic function

$$F(t) = I + X_1 t + X_2 t^2 + \dots$$

where the X_i are matrices in K_n . The series converges to a non-singular matrix, that is, to an element of G_K , for all t such that $v(t)$ is less than some fixed real number q . The expressions $f_i(t)$ denote the coordinates of $f(t)$, and the *tangent* (at I) of $f(t)$ is the matrix X_1 , as usual.

$f(t)$ is a *one-parameter-subgroup* if $f(t_1 + t_2) = f(t_1)f(t_2)$. The analytic curve $\exp tX$ for any matrix X is such a subgroup. The one-parameter subgroups here differ from those of ordinary Lie theory in the following respect: If $q > 0$ is any real number, then those values of t for which $v(t) \leq q$ and for which $f(t)$ is defined give values of $f(t)$ which form an entire group, not merely a local group. The fact that products exist in this group follows from the fact that v satisfies the condition (b).

From Theorems 1 and 2 it is seen that there exists a neighborhood of I in G_K in which for every element A , the matrix $X_A = \log A$ is defined and $\exp tX_A$ is a one-parameter subgroup which is defined for $v(t) \leq 1$ and passes through A for $t = 1$. All functions of the form $\exp tX$ are one-parameter subgroups in their region of convergence. A converse to this statement is contained in the following theorem.

THEOREM 4. *In G_K there exists a neighborhood of the identity in which there lies one and only one one-parameter subgroup with a given tangent.*

Using the uniqueness theorem for differential equations, (4), this can be proved exactly as in the theory of real Lie groups, (Pontrjagin, [15], pp. 185–187.) It follows that there exists a sphere S about I in which the only one-parameter subgroups are the exponential functions. Given any subgroup H of G_K and any sphere S_r contained in S we shall denote by H_r the subgroup $H \cap S_r$.

DEFINITION. *The Lie algebra L of a subgroup H of G_K is the set of tangents to analytic curves lying in some H_r . This is clearly independent of the choice of the number r .*

The commutator of two elements X, Y in the Lie algebra of G_K is defined as in Pontrjagin, [15], p. 238, and it can be shown that $[X, Y] = XY - YX$. The Lie algebra of G_K is then the Lie algebra K_{n1} , since any X in K_n is the tangent to an analytic curve $\exp tX$ in G_K . From the definition it follows as usual that the set L associated with a subgroup of G_K is a subalgebra of K_{n1} ; it also follows from the definition and Theorem 3 that it is an ideal (invariant subalgebra) if and only if H_r is an invariant subgroup.

6. Analytic Subgroups and Subalgebras

DEFINITION. Two subgroups H' and H'' of G_K are equivalent if there exists a sphere S_r around I such that $H'_r = H''_r$.

It can be shown that this is the same identification that is used in the theory of real local Lie groups, where H' and H'' are identified if their intersection is open in each. It is the object of this section to use this equivalence relation in obtaining a one-to-one correspondence between classes of equivalent subgroups of G_K and subalgebras of K_{n1} .

It has been shown that every subgroup of G_K has a Lie algebra which is a subalgebra of K_{n1} . Conversely, it will now be shown that if L is any subalgebra of K_{n1} , there is a closed subgroup of G_K whose Lie algebra is L . Let H be the totality of elements of G_K of the form $\exp X$, where X is an element of L , and such that $V(\exp X - I) < \rho^{1/(p-1)}$. By Theorem 1 we must have $V(X) < \rho^{1/(p-1)}$. If X and Y are in L and $V(X), V(Y) < \rho^{1/(p-1)}$, we have from Theorem 3 that $(\exp X)(\exp Y) = \exp Z$; here $V(\exp Z - I) < \rho^{1/(p-1)}$ and Z is in L because L is locally compact and Z is a limit of elements of L . H is therefore closed under multiplication. H must clearly contain the identity and the inverses of all its elements, since $\exp 0 = 1$ and $\exp(-X)$ is the inverse of $\exp X$. Hence H is a group. H is closed since it is a homeomorphic map of a bounded and closed subset of the locally compact space L .

DEFINITION. A subgroup H of G_K is analytic if it is equivalent to a subgroup constructed from a subalgebra as above.

THEOREM 5. If H is an analytic subgroup of G_K and $f(t)$ is any analytic curve in H with tangent a_1 , there exists in H a one-parameter subgroup with tangent a_1 .

PROOF. Let L be the subalgebra from which H has been constructed, and let S_r be a sphere contained in S . Let

$$f(t) = I + a_1 t + a_2 t^2 + \dots$$

be the given analytic curve. Let $f(t_1)$ be some point on this curve in H_r . We can then define

$$X_1 = \log f(t_1)$$

and the one-parameter subgroup

$$g_1(u) = \exp uX_1/t_1$$

passes through $f(t_1)$ for $u = t_1$ and lies in H_r for all values of u such that $V[g_1(u) - I] \leq r$. The matrix X_1 is in L since the group H is defined by L and $\exp X_1$ is in H .

The tangent to the curve $g_1(u)$ is

$$\begin{aligned} X'_1 &= X_1/t_1 = (1/t_1) \log f(t_1) \\ &= (1/t_1)[(a_1 t_1 + a_2 t_1^2 + \cdots) - (a_1 t_1 + a_2 t_1^2 + \cdots)^2/2 + \cdots] \\ &= a_1 + t_1 \epsilon \end{aligned}$$

where ϵ approaches zero with t_1 . Let us now consider a sequence of parameters t_1, t_2, \dots which approach zero. The corresponding one-parameter subgroups $g_1(u), g_2(u), \dots$ have tangents $X'_1 = a_1 + t_1 \epsilon, X'_2 = a_1 + t_2 \epsilon, \dots$

H is closed, and so for any fixed value of u , $\lim_{n \rightarrow \infty} g_n(u)$ is a point in H . The set of these points for all values of u under consideration can be shown to form a one-parameter subgroup $g(u)$ which is clearly not merely the identity. From Theorem 4, $g(u)$ must be an exponential function, so it is clear that its tangent must be the limit of the tangents X'_n , which is a_1 . The element a_1 must be in L , since L is locally compact. Q.E.D.

It follows from this theorem that if H is an analytic subgroup defined from a subalgebra L , then the Lie algebra of H is L . We thus obtain a one-to-one correspondence between the subalgebras of K_{n1} and the classes of equivalent analytic subgroups of G_K .

THEOREM 6. *Let H be an analytic subgroup of G_K with Lie algebra L of order m . There exists in G_K a system D of local analytic coordinates a''_i in a neighborhood S_r of I such that an element A of G_K is in H_r if and only if*

$$a''_i = 0, \quad i = m + 1, \dots, n^2.$$

The system D arises from the original system by an analytic transformation, and so by (2), we may say that H is defined by analytic functions.

PROOF. Let X be any matrix in K_{n1} such that $\exp X$ exists in G_K . We define the functions

$$h_i(X) = h_i(x_1, x_2, \dots, x_{n^2}) = (\exp X)_i, \quad i = 1, 2, \dots, n^2,$$

where the x_i are the elements of the matrix X in K_n and the $(\exp X)_i$ are the coordinates of this element in G_K . We have then $h_i(0, 0, \dots, 0) = 0$ and $(\partial h_i / \partial x_j)_0 = \delta_{ij}$, so the Jacobian of these functions is not zero at the origin. It follows from (3) that the equations

$$a_i = h_i(a'_1, a'_2, \dots, a'_{n^2}), \quad i = 1, 2, \dots, n^2,$$

can be solved for a'_i in terms of the original coordinates a_1, a_2, \dots, a_{n^2} . By (2), therefore, the numbers a'_i furnish a new local analytic coordinate system in G_K .

What we have done is to assign to each element A of G_K near I the elements of $\log A$ as its coordinates. Now given a subalgebra L of K_{n1} , we can change

the basis of K_n , so that a matrix of K_n is in L if and only if its last $n^2 - m$ coordinates are all zero. Such a transformation is algebraic, and analytic, and has a non-vanishing Jacobian at the origin. Relative to this basis, $\log A$ has n^2 new elements, and the last $n^2 - m$ of these are zero if and only if $\log A$ is in L . We assign these as coordinates in G_K and the conditions of the theorem are satisfied. Q.E.D.

It follows from this theorem that an analytic subgroup H of G_K has a local analytic coordinate system of dimension m and so is a p -adic Lie group.

THEOREM 7. *Let H be an analytic subgroup of G_K with Lie algebra L . Let X^1, X^2, \dots, X^m be a basis for L and $f_i(t)$ be analytic curves in H with tangents X^i respectively ($i = 1, 2, \dots, m$). Then there exists a neighborhood of I in H all of whose elements may be put in the form*

$$f_1(t_1) \cdot f_2(t_2) \cdots f_m(t_m).$$

PROOF. Let us define the functions

$$F_i(t_1, t_2, \dots, t_m) = [f_1(t_1) \cdot f_2(t_2) \cdots f_m(t_m)]_i, \quad i = 1, 2, \dots, n^2,$$

i denoting the coordinate of the expression in brackets relative to a coordinate system as described in the last theorem. These give an analytic mapping of a neighborhood of the origin of K^m into a neighborhood of I in H . From the independence of the tangents to the f_i , it can be shown that the Jacobian of these functions does not vanish at the origin. Hence we can solve these equations for the t_i in terms of the f_i and so by the last theorem the mapping defined by these functions covers an entire neighborhood of I in H . Q.E.D.

7. Analyticity of Subgroups

We have seen that an analytic subgroup H of G_K is closed and defined by analytic functions. It is now desirable to determine whether these two conditions are sufficient for analyticity. It will be shown that if $K = R_p$, any closed subgroup H is analytic and so is a p -adic Lie group. If $K \neq R_p$, however, this fact does not hold, as in the case of complex Lie groups. The subset of elements with coordinates in R_p is a closed subgroup but is obviously not analytic. We must therefore divide our argument into two parts.

Let H be a closed subgroup of G_K and let L be its Lie algebra. Let H' be an analytic subgroup corresponding to L and H'_τ, H_τ be intersections of H' and H respectively with some S_τ contained in S .

LEMMA. *Let A be any element in H_τ and let $X_A = \log A$. If $K = R_p$, H_τ contains $\exp tX_A$ for all t in K such that $v(t) \leq 1$. If $K \neq R_p$, but H is defined by analytic functions, the same holds.*

PROOF. i) Suppose $K = R_p$. If n is any rational integer,

$$\exp nX_A = (\exp X_A)^n$$

and so the expression on the left is in H_τ , since $\exp X_A$ is in H_τ . By (1), however, if t is any element of R_p such that $v(t) \leq 1$, it is a limit of a sequence of

rational integers t_m . Hence

$$\exp tX_A = \lim_{m \rightarrow \infty} \exp t_m X_A$$

and is in H_r since H is closed.

ii) Suppose $K \neq R_p$ but that H_r is defined by analytic functions, that is, there exist analytic functions $F_i(x_1, x_2, \dots, x_{n^2})$ such that a point A in G_K with coordinates $(a_1, a_2, \dots, a_{n^2})$ is in H_r if and only if $F_i(a_1, a_2, \dots, a_{n^2}) = 0$ for all i . By the above argument, $\exp tX_A$ is in H_r for all t in R_p such that $v(t) \leq 1$. Hence we have

$$F_i[(\exp tX_A)_1, (\exp tX_A)_2, \dots, (\exp tX_A)_{n^2}] = 0$$

for all i , and t in R_p . By (2) the F_i are analytic functions of t , and by (8) they must be zero for all t in K . Q.E.D.

THEOREM 8. *Let H satisfy the conditions in the lemma. Then H is analytic.*

PROOF. It follows from the lemma that H'_r contains H_r . By Theorem 7, a neighborhood of I in H'_r can be defined by analytic curves in H_r . This neighborhood is completely in H_r , so H is equivalent to the analytic subgroup H' . Q.E.D.

This theorem completes the setting up of the one-to-one correspondence between subgroups of G_K and subalgebras of K_{n^2} .

8. Some Special Groups

Before discussing the special groups, it will be necessary to prove a theorem which occurs in the ordinary theory of Lie groups.

THEOREM 9. *Let H be an analytic subgroup of G_K with subalgebra L . Let H_1 be an invariant subgroup of H with subalgebra L_1 , an ideal in L . Let \bar{H} be an analytic subgroup with Lie algebra \bar{L} and such that there exists a continuous homomorphism of H onto \bar{H} with H_1 being the set of elements mapped on the identity I . Then $\bar{L} \cong L/L_1$.*

PROOF: We must first show that any continuous homomorphic map of a one-parameter subgroup is an analytic curve. It is sufficient to prove that any continuous homomorphic map of the ring E_K of integers of K into G_K is analytic. We need only prove this for E_{R_p} since E_K is a direct product of a finite number of the E_{R_p} . Let $t \mapsto f(t)$ be such a mapping into some $S_r \subseteq S$ in G_K . We know that there exists an X such that

$$f(1) = \exp X,$$

and so

$$f(n) = \exp nX$$

for any rational integer n , since f is a homomorphic mapping. Since f is continuous, we have, by (1),

$$f(t) = \exp tX$$

for any t in E_{R_p} . This mapping is analytic.

It follows that if $f(t)$ is a one-parameter subgroup in H_r , its map is an analytic curve and one-parameter subgroup in \bar{H}_r and all one-parameter subgroups in \bar{H}_r are so obtained. This establishes a homomorphism between L and \bar{L} . Clearly, L_1 is the set mapped onto the zero element and so $\bar{L} = L/L_1$. Q.E.D.

Given a Lie algebra L in K_{n1} , we know that there is an infinite number of analytic subgroups of G_K whose Lie algebra is L . We are interested in finding, for a given L , an analytic subgroup H , defined by algebraic conditions, whose Lie algebra is L .

Let \mathfrak{A} be an associative algebra of order n over K and $D(\mathfrak{A})$ be the Lie algebra of derivations of \mathfrak{A} . Let H be the group of automorphisms of \mathfrak{A} . H and $D(\mathfrak{A})$ may be imbedded in G_K and K_{n1} respectively.

THEOREM 10. *The Lie algebra of H is $D(\mathfrak{A})$.*

PROOF. Let X be an element of $D(\mathfrak{A})$ such that $\exp X$ is defined. Let a and b be any two elements of \mathfrak{A} . Then

$$\begin{aligned} [(\exp tX) \cdot a][(\exp tX) \cdot b] &= \sum_{n=0}^{\infty} t^n \left[\sum_{p=0}^n (1/p!(n-p)!) (X^p a)(X^{n-p} b) \right] \\ &= \sum_{n=0}^{\infty} t^n / n! \left[\sum_{p=0}^n C_{n,p} (X^p a)(X^{n-p} b) \right] \\ &= \sum_{n=0}^{\infty} t^n / n! X^n(ab) \\ &= (\exp tX)ab. \end{aligned}$$

Hence $\exp tX$ is an automorphism in \mathfrak{A} for every t for which it is defined.

Conversely, let A be an element of H near I so that $A = \exp X$ for some matrix X . H is analytic, being algebraically defined, so $\exp tX$ is in H and hence is an automorphism in \mathfrak{A} for all values of t for which it is defined. We have, therefore,

$$(\exp tX) \cdot ab = [(\exp tX) \cdot a][(\exp tX) \cdot b].$$

Expanding these series in powers of t and multiplying out,

$$ab + (tX) \cdot ab + t\epsilon_1 = ab + ta(X \cdot b) + t(X \cdot a)b + t\epsilon_2$$

and

$$X \cdot ab + \epsilon_1 = a(X \cdot b) + (X \cdot a)b + \epsilon_2.$$

The quantities ϵ_1, ϵ_2 approach zero with t . Taking the limit of both sides as t approaches zero, we find X is a derivation. Q.E.D.

Let \mathfrak{A} be an associative algebra over K and J be an involution (involutorial anti-automorphism)³. An element a in \mathfrak{A} is called J -skew if $a^J = -a$. It is called J -orthogonal if $aa^J = a^J a = k \cdot 1$, where k is an element of K . The set \mathfrak{S} , of J -skew elements is a Lie algebra over K and the set \mathfrak{O} , of J -orthogonal elements is a multiplicative group.

³ For this step, see Jacobson, [10], p. 207.

⁴ For the required results on involutions, see Jacobson, [9], and Albert, [3], ch. X.

THEOREM 11. *The Lie algebra of \mathfrak{G}_J is $\mathfrak{S} = \mathfrak{S}_J \oplus K$.*

PROOF. The elements of \mathfrak{S} are in the form $z = a + k$, where a is in \mathfrak{S}_J and k is in K . When $\exp z$ exists, it is in \mathfrak{A} , since \mathfrak{A} has a finite basis and is closed. Hence $(\exp z)^J$ is defined. Any finite number of terms of the series $\exp z^J$ is equal to the corresponding number of terms of $(\exp z)^J$, so $\exp z^J$ exists and is equal to $(\exp z)^J$. Hence when $\exp(a + k)$ exists,

$$\begin{aligned} [\exp(a + k)][\exp(a + k)]^J &= [\exp(a + k)][\exp(-a + k)] \\ &= \exp 2k \in K, \end{aligned}$$

and so $\exp(a + k)$ is in \mathfrak{G}_J .

Conversely, if $(\exp tz)(\exp tz)^J = (\exp tz)^J(\exp tz) = k \in K$, for all t in some neighborhood of zero, then $zz^J = z^Jz$ and $(\exp z)(\exp z)^J = (\exp z)^J(\exp z) = \exp(z + z^J) = k$. If z is sufficiently small, $\log k$ exists and $z + z^J = \log k = k'$.

It is known that any z in \mathfrak{A} can be written uniquely as

$$z = b + d \quad \text{where} \quad b^J = -b, \quad d^J = d.$$

Hence $z + z^J = b + d + b^J + d^J = 2d = k'$. It follows that $d = k'/2$. Hence z must be in the form $a + k$ and so the Lie algebra of \mathfrak{G}_J is \mathfrak{S} . Q.E.D.

9. Groups of Simple Lie Algebras

A simple p -adic Lie group is defined as usual as a group with no invariant subgroup which is not discrete or equivalent to the whole group. These, of course, are the groups with the simple Lie algebras.

A simple Lie algebra over any field K of characteristic zero has been shown by Landherr, [13], to be normal simple over its extended center, which is a finite algebraic extension of K . We shall, therefore, restrict ourselves to the p -adic Lie groups whose Lie algebras are the "non-exceptional" normal simple Lie algebras over K . The normal simple associative algebras over a p -adic field K have been classified by Hasse, [6]. The Lie algebras which are normal simple over any field K of characteristic zero are shown by Jacobson, [12], to arise, except for the finite number of exceptional cases, from associative algebras over K in one of the following ways:

1) (Type A_I) Let \mathfrak{A} be a normal simple associative algebra over K , and let \mathfrak{A}_I be the Lie algebra obtained in the usual way. The derived algebra \mathfrak{A}'_I is normal simple.

2) (Types B, C, D) Let \mathfrak{A} be a normal simple associative algebra over K with an involution J of first kind. Then \mathfrak{S}_J is a normal simple Lie algebra.

3) (Type A_{II}) Let \mathfrak{A} be a simple associative algebra over K with center $\Sigma = K(q)$, where q^2 , but not q , is in K ; J is an involution of second kind. The derived algebra \mathfrak{S}'_J is a normal simple Lie algebra.

Using the results of Hasse, Jacobson has classified the normal simple Lie algebras over K and determined their automorphism groups. Any simple Lie algebra over K may be imbedded in $K_{n,1}$ and its automorphism group in G_K . The automorphism groups are analytic, being algebraically defined.

We wish to find for each normal simple Lie algebra L an analytic group H in G_K whose Lie algebra is L . Any other group in G_K with the same Lie algebra is equivalent to H .

We must consider separately the three cases:

1) *The group H of automorphisms of \mathfrak{A} has a Lie algebra isomorphic to \mathfrak{A}'_1 in Case 1.*

PROOF. Since \mathfrak{A} is the enveloping algebra of \mathfrak{A}_1 , and is simple, we have

$$\mathfrak{A}_1 = C \oplus \mathfrak{A}'_1$$

where C is abelian and hence must be the center of \mathfrak{A}_1 . (cf. Jacobson, [8].) C is then the center of \mathfrak{A} and so is isomorphic to K . Hence

$$\mathfrak{A}'_1 \cong \mathfrak{A}_1/K.$$

It is known, however, (Jacobson, [10]), that when \mathfrak{A} is normal simple,

$$D(\mathfrak{A}) \cong \mathfrak{A}_1/K.$$

By Theorem 10, therefore, the Lie algebra of H is isomorphic to \mathfrak{A}'_1 . Q.E.D.

2) *The group H of automorphisms of \mathfrak{S}_J has Lie algebra \mathfrak{S}_J .*

PROOF. Let K represent the abelian Lie algebra of scalar matrices. Let $\mathfrak{S} = \mathfrak{S}_J \oplus K$ and let \mathfrak{G}_J be the group of J -orthogonal elements of \mathfrak{A} .

We have shown that the Lie algebra of \mathfrak{G}_J is \mathfrak{S} . Now \mathfrak{G}_J is continuously homomorphic to H with K_0 being the set of elements mapped onto I . (cf. Jacobson, [9]; K_0 is the multiplicative group of K). It is easily seen that the Lie algebra of K_0 is K , so by Theorems 9 and 11, the Lie algebra of H is isomorphic to $\mathfrak{S}/K \cong \mathfrak{S}_J$. Q.E.D.

3) *Let H be the group of automorphisms of \mathfrak{S}'_J induced by inner automorphisms of \mathfrak{A} . The Lie algebra of H is \mathfrak{S}'_J .*

PROOF. The enveloping algebra of \mathfrak{S}_J is \mathfrak{A} . (cf. Jacobson, [11], p. 182). We have, therefore, as in Case 1),

$$\mathfrak{S}_J = C \oplus \mathfrak{S}'_J,$$

where C is the center of the Lie algebra \mathfrak{S}_J . The elements of C must commute with those of \mathfrak{S}_J in the associative multiplication, and hence with all of \mathfrak{A} , so

$$C = \mathfrak{S}_J \cap \Sigma,$$

since Σ is the center of \mathfrak{A} .

We have proved that the Lie algebra of \mathfrak{G}_J is $\mathfrak{S} = \mathfrak{S}_J \oplus K$, and we have

$$\mathfrak{S} = \mathfrak{S}'_J \oplus (\mathfrak{S}_J \cap \Sigma) \oplus K.$$

It is known (Albert, [3]) that

$$\Sigma = K \oplus (\mathfrak{S}_J \cap \Sigma).$$

We have, therefore, $\mathfrak{S} = \mathfrak{S}'_J \oplus \Sigma$. It has been shown by Jacobson, [11], (p. 185), that \mathfrak{G}_J is homomorphic to H . This can easily be seen to be continuous

and the group Σ_0 is the group mapped on the identity. It follows that H has Lie algebra \mathfrak{S}'_j . Q.E.D.

Although we have treated these cases separately, the result is the same in each case. Let L be the Lie algebra arising from \mathfrak{A} in one of these three ways. From the results of Jacobson, [12], (pp. 339, 340), and from the fact that all automorphisms of a normal simple algebra are inner, the group H of automorphisms in L induced by inner automorphisms in \mathfrak{A} has Lie algebra L in each case.

PRINCETON UNIVERSITY

BIBLIOGRAPHY

1. ADO, I. Bull. de la Soc. Physico-Math. de Kazan, **6** (1935).
2. ALBERT, A. *Modern Higher Algebra*, U. of Chicago Press, 1937.
3. ALBERT, A. *Structure of Algebras*, Am. Math. Soc. Colloquium Publications, vol. XXIV, New York, 1939.
4. CHABAUTY, C. *Sur les équations diophantiennes*, etc., *Annali di Matematica*, **17** (1938), pp. 127-168.
5. CHEVALLEY, C. *Sur la théorie du corps de classes*, etc., *Journal of the Faculty of Science, Tokyo*, **2** (1933), pp. 365-474.
6. HASSE, H. *Über p -adische Schiefkörper*, *Math. Annalen*, **104** (1931), pp. 495-534.
7. HAUSDORFF, F. *Die symbolische Exponentialformel*, etc., *Ber. der Ges. der Wiss. zu Leipzig*, **58** (1906), pp. 19-48.
8. JACOBSON, N. *Rational Methods in the Theory of Lie Algebras*, these *Annals*, **36** (1935), pp. 875-881.
9. JACOBSON, N. *A Class of Normal Simple Lie Algebras of Characteristic Zero*, these *Annals*, **38** (1937), pp. 508-517.
10. JACOBSON, N. *Abstract Derivation and Lie Algebras*, *Trans. Am. Math. Soc.*, **42** (1937), pp. 206-224.
11. JACOBSON, N. *Simple Lie Algebras of Type A*, these *Annals*, **39** (1938), pp. 181-188.
12. JACOBSON, N. *Simple Lie Algebras over a Field of Characteristic Zero*, *Duke Math. J.*, **4** (1938), pp. 534-551.
13. LANDHERR, W. *Über einfache Liesche Ringe*, *Hamb. Abhandlungen*, **11** (1935), pp. 41-64.
14. LUTZ, E. *Sur l'équation $y^2 = x^3 - Ax - B$* , etc., *J. für die reine und angew. Math.*, **177** (1937), pp. 238-247.
15. PONTRJAGIN, L. *Topological Groups*, Princeton, 1939.

ON THE MODULAR REPRESENTATIONS OF THE SYMMETRIC GROUP

By R. M. THRALL AND C. J. NESBITT

(Received December 11, 1941)

1. Introduction

The purpose of the present paper is to determine all modular (matrix) representations of the symmetric group, \mathfrak{S}_m , of degree m , where $m < 2p$. The elements of the representing matrices are to be chosen from any field, \mathfrak{f} , of characteristic p . Every indecomposable (modular) representation of \mathfrak{S}_m is equivalent to a rational one (i.e. to one in which \mathfrak{f} is the prime field) so the nature of the field, \mathfrak{f} , is of no particular interest in what follows.

In the last several years considerable progress has been made in the theory of modular representations of finite groups.¹ This theory is especially well worked out in case the order of the group is divisible by only the first power of p ; hence our requirement $m < 2p$. (Actually we treat here only the cases $p \leq m < 2p$ since for $m < p$ the ordinary theory applies, leaving no problem.)

Any representation of a group (or of an algebra) is completely characterized by its indecomposable constituents. In general there are an infinite number of inequivalent indecomposable modular representations of a finite group. One of the main results of the present paper is a proof that the symmetric group of degree less than $2p$ has only a finite number of inequivalent indecomposable representations. In sections 2-4 we determine the structure of the regular representation of \mathfrak{S}_m (or of its \mathfrak{f} group ring \mathfrak{R}_m). In section 5 we show how any indecomposable representation can be built up from "elementary modules" (see section 3 for definition and references) of the group ring.

In section 6 we state Nakayama's results² on the modular representations of \mathfrak{S}_m and add a discussion of the behavior of representations of \mathfrak{S}_m when considered only for elements of \mathfrak{S}_{m-1} .

The final three sections are devoted to specific determination of all representations of \mathfrak{S}_p with indications of generalization to \mathfrak{S}_{p+1} , \mathfrak{S}_{p+2} .

2. Preliminaries

Let $m = p + l$, $l < p$, and let

$$(1) \quad m = \alpha_1 + 2\alpha_2 + \cdots + m\alpha_m$$

be a partition of m . Corresponding to this partition (α) there is a class $C(\alpha)$ of conjugate elements of \mathfrak{S}_m such that if $s \in C(\alpha)$, then s is a product of α_1 1-cycles,

¹ See the bibliography at the end of the paper.

² See [8] especially part II.

α_2 2-cycles, \dots , α_m m -cycles. The number of conjugate classes, and hence the number of ordinary irreducible representations is P_m , the number of partitions of m .

Since $m = p + l$, α_p in (1) may be 0 or 1. The class $C(\alpha)$ is p -singular (that is, the order of the elements of $C(\alpha)$ is divisible by p) if and only if $\alpha_p = 1$. Then the number of p -singular classes is equal to P_l . It follows that the number of modular irreducible representations which is equal to the number of p -regular classes is $P_m - P_l$.³

Corresponding to the decomposition of the group ring \mathfrak{R}_m into a direct sum of directly indecomposable invariant subalgebras there exists a classification of the ordinary irreducible and the modular irreducible representations, and their characters into blocks.⁴ A block \mathfrak{B}_r is said to be of type β if all the ordinary irreducible representations which belong to \mathfrak{B}_r have degrees $\equiv 0 \pmod{p^\beta}$, but at least one of these degrees $\not\equiv 0 \pmod{p^{\beta+1}}$. In our case we have just blocks of type 0 or lowest kind, and blocks of highest kind (here of type 1).

Theorem II of [1] states that the number t_0 of blocks of lowest kind is equal to the number of p -regular classes of conjugate elements where the number of elements in the class is prime to p . The number of elements in the class $C(\alpha)$ is $m!/n(\alpha)$ where $n(\alpha) = \alpha_1! \alpha_2! 2^{\alpha_2} \dots \alpha_m! m^{\alpha_m}$ is the order of the normalizer of any element s in $C(\alpha)$. To determine t_0 for our case, we count the partitions of m where $\alpha_p = 0$ (that is, $C(\alpha)$ is p -regular) and $m!/n(\alpha) \not\equiv 0 \pmod{p}$. Since for $i > 1$, α_i is less than p , and α_p is here 0, then $n(\alpha) \equiv 0 \pmod{p}$ only if $\alpha_1 \geq p$. One easily sees a 1-1 correspondence between the partitions of m with $\alpha_1 \geq p$, and the partitions of l . We thus obtain $t_0 = P_l$.

We denote by x_r , y_r the numbers of ordinary and of modular irreducible characters which belong to a block \mathfrak{B}_r . If \mathfrak{B}_r is of highest kind $x_r = y_r = 1$; for \mathfrak{B}_r of lowest kind $x_r \geq y_r + 1$.⁵ But, by the above, $\sum x_r - \sum y_r = P_m - (P_m - P_l) = P_l$, and there are P_l blocks of lowest kind, so the only possibility is that for each block of lowest kind $x_r = y_r + 1$. We shall show below that $x_r = p$ for blocks of lowest kind.

In the theory of modular representations of groups two sets of numbers play leading roles: the decomposition numbers describe the splitting of the ordinary irreducible representations (when taken in the modular sense) into modular irreducible constituents; the Cartan invariants give the multiplicities of the modular irreducible representations as constituents of the indecomposable parts of the modular regular representation. If D_r , C_r denote the matrices of decomposition and Cartan numbers for the block B_r , then a main theorem is that⁶

$$(2) \quad C_r = D_r' D_r.$$

³ Cf. Brauer-Nesbitt [1], §8 for proof that the number of p -regular classes is equal the number of modular irreducible representations.

⁴ Brauer-Nesbitt [1], §9, and Nakayama [9], Theorem 5.

⁵ Brauer-Nesbitt [1], §19, Theorem 5.

⁶ Brauer-Nesbitt [1], §§4, 5 and 9.

From Brauer's work (in particular, see Theorem 14 of [2]) we have in our case that for a block of lowest kind

$$(3) \quad D_r = \begin{vmatrix} 1 & & & & \\ & 11 & & & \\ & & 11 & & \\ & & & \dots\dots & \\ & & & & 11 \\ & & & & & 1 \end{vmatrix}$$

Here D_r has x_r rows, y_r columns. Then C_r has y_r rows, y_r columns, and from (2), (3)

$$(4) \quad C_r = \begin{vmatrix} 21 & & & & \\ & 121 & & & \\ & & 121 & & \\ & & & \dots\dots & \\ & & & & 121 \\ & & & & & 12 \end{vmatrix}$$

It follows from (4) that $\det C_r$ is $y_r + 1 = x_r$.

We consider now the set M of $n(\alpha)$'s corresponding to the p -regular classes $C(\alpha)$, and form the product $\prod n(\alpha)$. For each of the P_i partitions (α) with $\alpha_p = 0$, $\alpha_1 \geq p$, $n(\alpha)$ is divisible by p , and all other $n(\alpha)$ of the set $M \not\equiv 0 \pmod{p}$. Then p^{P_i} is the highest power of p which divides $\prod n(\alpha)$. By Theorem I of [3] the determinant of the complete matrix C of Cartan invariants is then equal to p^{P_i} , that is,

$$\det C = \prod \det C_r = p^{P_i}.$$

But for blocks B_r of highest kind $\det C_r = 1$, and for blocks of lowest kind $\det C_r = x_r$, and there are P_i blocks of lowest kind; hence for each such block $x_r = p$. *To each block of lowest kind there belong p ordinary irreducible representations and $p - 1$ modular irreducible representations.*

It follows also that there are $P_m - pP_i$ blocks of highest kind and that the total number of blocks is $P_m - (p - 1)P_i$.

3. Loewy form

Let \mathfrak{A} denote any matrix representation of \mathfrak{N}_m . We consider the algebra \mathfrak{A} as a system of linear transformations of a vector space \mathfrak{B} of suitable dimension. Let $\mathfrak{N}, \mathfrak{N}^2, \dots, \mathfrak{N}^{i-1}, \mathfrak{N}^i = (0)$ denote the powers of the radical of \mathfrak{A} . We form the upper Loewy series of \mathfrak{B} , namely⁷

$$(5) \quad \mathfrak{B} \supset \mathfrak{N}\mathfrak{B} \supset \mathfrak{N}^2\mathfrak{B} \dots \supset \mathfrak{N}^{i-1}\mathfrak{B} \supset 0.$$

⁷ For a discussion of Loewy series see §§2, 5 of [4]. That (5) is the upper series for \mathfrak{B} follows by proving Theorem 12.3A of [4] for vector spaces having \mathfrak{A} as operator system rather than for ideals of \mathfrak{A} .

Here $\mathfrak{N}^{p-1}\mathfrak{B}/\mathfrak{N}^p\mathfrak{B}$ is the unique maximal completely reducible factor group that may be obtained from $\mathfrak{N}^{p-1}\mathfrak{B}$ ($p = 1, 2, \dots, t$). If we adapt the coordinate system in \mathfrak{B} to the series (5), then \mathfrak{A} appears in upper Loewy form,

$$(6) \quad \mathfrak{A} \sim \left\| \begin{array}{c} \mathfrak{L}_1(\mathfrak{A}) \\ \mathfrak{L}_2(\mathfrak{A}) \\ * \\ \mathfrak{L}_t(\mathfrak{A}) \end{array} \right\|$$

where the $\mathfrak{L}_i(\mathfrak{A})$ are the upper Loewy constituents of \mathfrak{A} , and are completely reducible. t may be called the Loewy length of \mathfrak{A} .⁸

We consider now the Loewy form of the indecomposable parts of the regular representation \mathfrak{R} of \mathfrak{R}_m . Let $\mathfrak{F}_1, \mathfrak{F}_2, \dots, \mathfrak{F}_k$ denote the modular irreducible representations of \mathfrak{R}_m . It follows from [6] (see, in particular, Theorems 2, 3, and 8) that we may denote the indecomposable parts of \mathfrak{R} by $\mathfrak{U}_1, \dots, \mathfrak{U}_k$ where, if \mathfrak{U}_κ is written in the form (6), $\mathfrak{L}_1(\mathfrak{U}_\kappa) = \mathfrak{L}_t(\mathfrak{U}_\kappa) = \mathfrak{F}_\kappa$

$$(7) \quad \mathfrak{U}_\kappa = \left\| \begin{array}{c} \mathfrak{F}_\kappa \\ \mathfrak{L}_2(\mathfrak{U}_\kappa) \\ \cdot \\ * \quad \cdot \\ \mathfrak{L}_{t-1}(\mathfrak{U}_\kappa) \\ \mathfrak{F}_\kappa \end{array} \right\|.$$

We have considered here only the upper Loewy forms. We might similarly have discussed the lower Loewy forms. For the indecomposable parts \mathfrak{U}_κ of \mathfrak{R}_m (see below) the upper Loewy forms coincide with the lower Loewy forms.

In the following we shall use the notation $c_{\lambda\kappa}$ to denote the multiplicity of \mathfrak{F}_λ as constituent of \mathfrak{U}_κ ; $c_{\lambda\kappa}$ is a Cartan invariant.

4. Elementary modules

Let \mathfrak{R} denote the modular regular representation of the group ring \mathfrak{R}_m of \mathfrak{S}_m . It is well known that \mathfrak{R} is a faithful representation of \mathfrak{R}_m . Then instead of considering elements of \mathfrak{R}_m we shall for the time being speak of the corresponding elements of \mathfrak{R} . We assume \mathfrak{R} to be in reduced form, that is

$$\mathfrak{R} = \left\| \begin{array}{ccc} \mathfrak{R}_{11} & & \\ \mathfrak{R}_{21} & \mathfrak{R}_{22} & \\ \vdots & \ddots & \\ \mathfrak{R}_{s1} & \dots & \mathfrak{R}_{ss} \end{array} \right\|,$$

where the \mathfrak{R}_{ij} denote irreducible constituents of \mathfrak{R} . We may further assume that $\mathfrak{R} = \mathfrak{R}^* + \mathfrak{R}$, where \mathfrak{R}^* is the semi-simple algebra obtained from \mathfrak{R} by replacing the \mathfrak{R}_{ij} with $i > j$ by 0, and where \mathfrak{R} , the radical of \mathfrak{R} , has 0 in place

⁸ Cf. Brauer, [4], §5.

of the \mathfrak{R}_{ii} in the main diagonal. Let, as before, $\mathfrak{F}_1, \mathfrak{F}_2, \dots, \mathfrak{F}_k$ denote the distinct modular irreducible representations of \mathfrak{R}_m , and f_1, f_2, \dots, f_k their degrees. We mean by \mathfrak{R}_κ^* the simple subalgebra of \mathfrak{R}^* which has 0 in place of all \mathfrak{R}_{ii} except those which are equivalent to \mathfrak{F}_κ . Let $e_\kappa(ij)$ ($i, j = 1, 2, \dots, f_\kappa$) be a set of matrix units for \mathfrak{R}_κ^* . We denote the unit element of the simple algebra \mathfrak{R}_κ^* by $e_\kappa = \sum_{i=1}^{f_\kappa} e_\kappa(ii)$. An element a of \mathfrak{R} we say is of type (κ, λ) if $e_\kappa a e_\lambda = a$. In [6] it was shown that a system of *primitive* elements

$$(8) \quad b_1, b_2, \dots, b_m, \quad m = \sum_{\kappa, \lambda=1}^k c_{\kappa\lambda}$$

could be chosen so that these and their right and left products with suitable $e_\kappa(ij)$ gave a basis for \mathfrak{R} . More fully, for each type (κ, λ) there are $c_{\kappa\lambda}$ elements b in the set (8) of that type. If b_ρ is of type (κ, λ) then we take all the elements

$$(9) \quad e_\kappa(i1)b_\rho e_\lambda(1j) \quad (i = 1, 2, \dots, f_\kappa, \quad j = 1, 2, \dots, f_\lambda)$$

for part of our basis of \mathfrak{R} . Doing this for each b_ρ we obtain what has been called the Cartan basis of \mathfrak{R} . This Cartan basis can be so arranged that the regular representation with respect to this new basis is split into indecomposable and irreducible parts.

The Cartan basis system is the starting point for the definition recently given by W. M. Scott of *elementary modules*.⁹ An element a of \mathfrak{R} , expressed in terms of the Cartan basis elements will have the form

$$a = \sum_{\rho, i, j} h_{ij}^\rho(a) e_\kappa(i1) b_\rho e_\lambda(1j).$$

Scott arranges the coefficients $h_{ij}^\rho(a)$, for a fixed ρ , in a matrix $H_\rho(a) = ||h_{ij}^\rho(a)||$. The Abelian additive group generated by the matrices $H^\rho(a)$, $a \in \mathfrak{R}$, Scott has called an elementary module. The significance of these for us here is that the representations of \mathfrak{R}_m , in particular, the regular representations, may be expressed in simple form by these elementary modules.

We have two cases to consider:

(1) *Blocks of highest kind.* Let \mathfrak{B}_τ be a block of highest kind and \mathfrak{F}_λ be the unique modular irreducible representation belonging to \mathfrak{B}_τ . Then $c_{\lambda\mu} = 0$ for $\lambda \neq \mu$, $c_{\lambda\lambda} = 1$, and the elements $e_\lambda(ij)$ ($i, j = 1, 2, \dots, f_\lambda$) may be taken as the Cartan basis for \mathfrak{B}_τ . The corresponding elementary module is just \mathfrak{F}_λ . Further, $\mathfrak{F}_\lambda = \mathfrak{U}_\lambda$, where \mathfrak{U}_λ is the unique indecomposable part of \mathfrak{R} corresponding to \mathfrak{F}_λ .

(2) *Blocks of lowest kind.* Let \mathfrak{B}_τ be a block of lowest kind, and let us choose our enumeration of the modular irreducible representations so that $\mathfrak{F}_1, \mathfrak{F}_2, \dots, \mathfrak{F}_{p-1}$ belong to \mathfrak{B}_τ , and further so that the matrix C_τ of Cartan invariants for \mathfrak{B}_τ is in the form (4).

Then for $\kappa \neq 1$ or $p-1$, $c_{\kappa\kappa} = 2$, $c_{\kappa-1, \kappa} = 1 = c_{\kappa+1, \kappa}$, and all other $c_{\lambda\kappa}$ are zero. Corresponding to these invariants there exist primitive elements $e_\kappa(11)$ and $b_{\kappa\kappa}$ of type (κ, κ) , $b_{\kappa-1, \kappa}$ of type $(\kappa-1, \kappa)$ and $b_{\kappa+1, \kappa}$ of type $(\kappa+1, \kappa)$.

⁹ Scott, [10].

The Cartan basis elements

$$\begin{array}{lll}
 \text{(a)} & e_{\kappa}(j1) & j = 1, 2, \dots, f_{\kappa} \\
 \text{(10) (b)} & e_{\kappa-1}(k1)b_{\kappa-1,\kappa} & k = 1, 2, \dots, f_{\kappa-1} \\
 & e_{\kappa+1}(l1)b_{\kappa+1,\kappa} & l = 1, 2, \dots, f_{\kappa+1} \\
 \text{(c)} & e_{\kappa}(m1)b_{\kappa\kappa} & m = 1, 2, \dots, f_{\kappa}
 \end{array}$$

form a basis for an indecomposable \mathfrak{R} -left ideal $\mathfrak{l}_{\kappa} = \mathfrak{R}e_{\kappa}(11)$ which is a summand in the expression of \mathfrak{R} as a direct sum of indecomposable left ideals. \mathfrak{l}_{κ} may be considered as the representation space of the indecomposable part \mathfrak{U}_{κ} of \mathfrak{R} . The form (7) of \mathfrak{U}_{κ} shows that the Loewy length t of $\mathfrak{U}_{\kappa} \geq 3$. It cannot be greater than 3 for then $\mathfrak{N}^3\mathfrak{l}_{\kappa} \neq 0$, that is, there would have to exist primitive elements of type $(*, \kappa)$ belonging to \mathfrak{N}^3 . Then t must equal 3, and again from (7) it follows that $\mathfrak{N}^2\mathfrak{l}_{\kappa}$ is generated by the elements (c), $b_{\kappa\kappa}$ belongs to \mathfrak{N}^2 , and that $b_{\kappa-1,\kappa}, b_{\kappa+1,\kappa}$ belong to \mathfrak{N} . We may assume the primitive elements so chosen that $b_{\kappa,\kappa-1}b_{\kappa-1,\kappa} = b_{\kappa,\kappa+1}b_{\kappa+1,\kappa} = b_{\kappa,\kappa}$ (here $b_{\kappa,\kappa-1}, b_{\kappa,\kappa+1}$ correspond to the Cartan invariants $c_{\kappa,\kappa-1} = c_{\kappa,\kappa+1} = 1$). We denote by $\mathfrak{F}_{\kappa}, \mathfrak{S}_{\kappa}^{\kappa-1}, \mathfrak{S}_{\kappa}^{\kappa+1}, \mathfrak{S}_{\kappa}^{\kappa}$ the elementary modules corresponding to $e_{\kappa}(11), b_{\kappa-1,\kappa}, b_{\kappa+1,\kappa}$ and $b_{\kappa,\kappa}$. Then following the method of [6], §3, taking in our basis of \mathfrak{l}_{κ} first the elements (a), then those of (b) and lastly those of (c) we calculate \mathfrak{U}_{κ} to be

$$(11) \quad \mathfrak{U}_{\kappa} = \left\| \begin{array}{cccc} \mathfrak{F}_{\kappa} & & & \\ \mathfrak{S}_{\kappa}^{\kappa-1} & \mathfrak{F}_{\kappa-1} & & \\ \mathfrak{S}_{\kappa}^{\kappa+1} & 0 & \mathfrak{F}_{\kappa+1} & \\ \mathfrak{S}_{\kappa}^{\kappa} & \mathfrak{S}_{\kappa-1}^{\kappa} & \mathfrak{S}_{\kappa+1}^{\kappa} & \mathfrak{F}_{\kappa} \end{array} \right\|$$

We similarly can compute

$$\begin{aligned}
 \mathfrak{U}_1 &= \left\| \begin{array}{ccc} \mathfrak{F}_1 & & \\ \mathfrak{S}_1^2 & \mathfrak{F}_2 & \\ \mathfrak{S}_1^1 & \mathfrak{S}_2^1 & \mathfrak{F}_1 \end{array} \right\| \\
 \mathfrak{U}_{p-1} &= \left\| \begin{array}{ccc} \mathfrak{F}_{p-2} & & \\ \mathfrak{S}_{p-1}^{p-2} & \mathfrak{F}_{p-2} & \\ \mathfrak{S}_{p-1}^{p-1} & \mathfrak{S}_{p-2}^{p-1} & \mathfrak{F}_{p-1} \end{array} \right\|
 \end{aligned}$$

5. Indecomposable representations of \mathfrak{R}_m

We set out now to determine all indecomposable representations of \mathfrak{R}_m . A first clue is that the Loewy length $l(\mathfrak{M})$ of any representation \mathfrak{M} of $\mathfrak{R}_m \leq 3$. This follows from our above result that $l(\mathfrak{U}_{\kappa}) = 1$ or 3 according as \mathfrak{U}_{κ} belongs to a block of highest or of lowest kind, and Theorems 6.6C, 11.5B of [3].

A second simplification comes from observing that \mathfrak{M} may be taken in upper

Loewy form and at the same time have its simple parts expressed in terms of the elementary modules \mathfrak{S} . Let us picture \mathfrak{M} in reduced form

$$\mathfrak{M} = \left\| \begin{array}{cc} \mathfrak{M}_{11} & \\ \mathfrak{M}_{21} & \mathfrak{M}_{22} \\ \dots\dots\dots \end{array} \right\|$$

and denote by \mathfrak{M}^* the subalgebra of \mathfrak{M} obtained by replacing in \mathfrak{M} the parts \mathfrak{M}_{ij} , $j < i$ by 0. It follows from Scott's results,¹⁰ that if in the representation \mathfrak{M} , the semisimple algebra \mathfrak{R}^* of \mathfrak{R} is mapped into \mathfrak{M}^* , then the simple parts \mathfrak{M}_{ij} are expressible as linear combinations of the elementary modules \mathfrak{S} . Let \mathfrak{B} be the representation space of \mathfrak{M} , and suppose the Loewy length of \mathfrak{M} is 3. We take the Loewy series $\mathfrak{B} \supset \mathfrak{N}\mathfrak{B} \supset \mathfrak{N}^2\mathfrak{B} \supset 0$. Here $\mathfrak{N}^2\mathfrak{B}$ may be considered as an \mathfrak{R}^* space, and as such is a direct sum of irreducible \mathfrak{R}^* spaces, $\mathfrak{N}^2\mathfrak{B} = \mathfrak{B}_1^{(2)} \oplus \dots \oplus \mathfrak{B}_{\alpha_2}^{(2)}$. Further $\mathfrak{N}\mathfrak{B}$ as an \mathfrak{R}^* space is a direct sum of $\mathfrak{N}^2\mathfrak{B}$ and a complementary space which may also be written as a direct sum of irreducible \mathfrak{R}^* spaces. By continuing the argument, we obtain that as an \mathfrak{R}^* space

$$\mathfrak{B} = \mathfrak{B}_1^{(0)} \oplus \dots \oplus \mathfrak{B}_{\alpha_0}^{(0)} \oplus \mathfrak{B}_1^{(1)} \oplus \dots \oplus \mathfrak{B}_{\alpha_1}^{(1)} \oplus \mathfrak{B}_1^{(2)} \oplus \dots \oplus \mathfrak{B}_{\alpha_2}^{(2)}.$$

Adapting the co-ordinate system to this decomposition of \mathfrak{B} , we obtain \mathfrak{R}^* mapped on \mathfrak{M}^* in \mathfrak{M} , and simultaneously have \mathfrak{M} in Loewy form.

Let now \mathfrak{M} be a modular indecomposable representation of \mathfrak{R}_m and let $\mathfrak{R}_m = \mathfrak{I}_1 \oplus \mathfrak{I}_2 \oplus \dots \oplus \mathfrak{I}_q$ denote the decomposition of \mathfrak{R}_m into indecomposable two-sided ideals. Since the representation space may be written

$$\mathfrak{B} = \mathfrak{R}_m\mathfrak{B} = \mathfrak{I}_1\mathfrak{B} \oplus \mathfrak{I}_2\mathfrak{B} \oplus \dots \oplus \mathfrak{I}_q\mathfrak{B}$$

and \mathfrak{M} is indecomposable, we find that only one $\mathfrak{I}_i\mathfrak{B}$, say $\mathfrak{I}_i\mathfrak{B}$, can be different from zero. This implies that only \mathfrak{I}_i is mapped on something different from zero in the representation \mathfrak{M} , and that \mathfrak{M} contains only modular irreducible representations belonging to the block corresponding to \mathfrak{I}_i .

We use the Loewy length $l(\mathfrak{M})$ of \mathfrak{M} to distinguish three cases.

1) $l(\mathfrak{M}) = 1$. Then \mathfrak{M} is both completely reducible and indecomposable, and so \mathfrak{M} must be equivalent to a modular irreducible representation \mathfrak{F}_α . We observe that if \mathfrak{M} contains a modular irreducible constituent belonging to a block of highest kind then $l(\mathfrak{M}) = 1$.¹¹

2) $l(\mathfrak{M}) = 2$. Then the modular irreducible constituents of \mathfrak{M} must belong to a block \mathfrak{B}_r of lowest kind. Let $\mathfrak{F}_1, \mathfrak{S}_1^1, \mathfrak{S}_2^1, \mathfrak{S}_1^2, \mathfrak{F}_2, \mathfrak{S}_2^2, \mathfrak{S}_3^2, \dots, \mathfrak{S}_{p-2}^{p-1}, \mathfrak{F}_{p-1}, \mathfrak{S}_{p-1}^{p-1}$ denote the elementary modules of \mathfrak{B}_r ; here $\mathfrak{F}_1, \dots, \mathfrak{F}_{p-1}$ are the modular irreducible representations of \mathfrak{B}_r , $\mathfrak{S}_\lambda^\kappa$, $\lambda \neq \kappa$, are the elementary modules of \mathfrak{B}_r which belong to the first power \mathfrak{N} of the radical, and the \mathfrak{S}_i^κ are those which belong to \mathfrak{N}^2 . As here the Loewy series for \mathfrak{M} is $\mathfrak{B} \supset \mathfrak{N}\mathfrak{B} \supset (0)$,

¹⁰ Cf. Scott, [10].

¹¹ For a modular irreducible representation belonging to a block of highest kind is also an indecomposable part of the regular representation. Then apply Remark 3 of [7].

the \mathfrak{F}_κ^* will not appear in \mathfrak{M} . Our notation for the elementary modules is based on the form (4) of the matrix C , of Cartan invariants. From the above considerations, we may suppose that \mathfrak{M} in Loewy form is written (denoting the unit matrix of degree f by E_f)

$$(12) \quad \mathfrak{M} = \begin{vmatrix} E_{h_1} \times \mathfrak{F}_1 & & & & \\ & E_{h_3} \times \mathfrak{F}_3 & & & \\ & & \ddots & & \\ & & & E_{h_{p-2}} \times \mathfrak{F}_{p-2} & \\ A_1^2 \times \mathfrak{F}_1^2, & A_3^2 \times \mathfrak{F}_3^2 & & & E_{h_2} \times \mathfrak{F}_2 \\ & A_3^4 \times \mathfrak{F}_3^4, & & & & E_{h_4} \times \mathfrak{F}_4 \\ & & \ddots & & & & \ddots \\ & & & A_{p-2}^{p-1} \times \mathfrak{F}_{p-2}^{p-1} & & & & E_{h_{p-1}} \times \mathfrak{F}_{p-1} \end{vmatrix}$$

or in a similar form with the even \mathfrak{F}_κ in the top Loewy constituent and the odd \mathfrak{F}_κ in the bottom constituent. If we assumed that both even and odd constituents \mathfrak{F}_κ appear in the same Loewy constituent, then by permutation of the rows and columns in \mathfrak{M} a decomposition of \mathfrak{M} would be obtained.

From Schur's lemma it follows that a matrix P which commutes with \mathfrak{M} has the form

$$(13) \quad P = \begin{vmatrix} P_1 \times E_{h_1} & & & \\ & P_3 \times E_{h_3} & & \\ & & \ddots & \\ & & & P_{p-1} \times E_{h_{p-1}} \end{vmatrix}$$

where P_i is a square matrix of h_i rows. In addition in order that P commute with \mathfrak{M} the following relations must be satisfied

$$(14) \quad \begin{aligned} A_1^2 P_1 &= P_2 A_1^2 \\ A_3^2 P_3 &= P_2 A_3^2 \\ &\dots\dots\dots \\ A_{p-2}^{p-1} P_{p-2} &= P_{p-1} A_{p-2}^{p-1}. \end{aligned}$$

Here A_q^p has h_p rows, h_q columns.

We assume first that all h_i ($i = 1, 2, \dots, p-1$) are different from 0. Then the relations (14), together with the theorem¹² that a matrix P commuting with an indecomposable representation \mathfrak{M} can have just one distinct characteristic root, are sufficient to show that each $h_i = 1$.

In outline the argument is this. We take P with $P_2 = P_3 = \dots = P_{p-1} = 0$, then the only relation (14) remaining is

$$(15) \quad A_1^2 P_1 = 0.$$

¹² Cf. Brauer-Schur, [11].

If now the rank of A_1^2 were less than h_1 we could find a P_1 with characteristic root $\neq 0$ to satisfy (15), contrary to the Schur-Brauer theorem. Then the rank of $A_1^2 = h_1$, and so $h_2 \geq h_1$. Now taking $P_3 = P_4 = \dots = P_{p-1} = 0$, and choosing P_1 to satisfy $A_1^2 P_1 = P_2 A_1^2$, the remaining relation (14) is $P_2 A_3^2 = 0$ and by the same argument as before the rank of A_3^2 is h_2 , and $h_3 \geq h_2$. Next, we take $P_4 = P_5 = \dots = P_{p-1} = 0$, choose P_1, P_2 to satisfy the first two relations (14), and consider $A_3^2 P_3 = 0$. We obtain in this manner that

$$h_1 \leq h_2 \leq h_3 \leq \dots \leq h_{p-1}.$$

If, however, we had started at the other end of the relations (14), we would obtain

$$h_{p-1} \leq h_{p-2} \leq \dots \leq h_1$$

so that $h_1 = h_2 = \dots = h_{p-1} = c$, say, and further the A_q^2 are all of rank c and so are non-singular. It follows that the relations (14) are satisfied when P_1 is arbitrary, $P_2 = (A_1^2)^{-1} P_1 A_1^2$, $P_3 = (A_3^2)^{-1} P_2 A_3^2$, etc. Then P_1 must be of degree 1 (for otherwise P_1 could have two different characteristic roots) and so $c = 1$.

The second step in this argument requires some elaboration. Since we have seen A_1^2 is of rank h_1 , we may find non-singular matrices X and Y such that $XA_1^2Y = \begin{vmatrix} E_{h_1} \\ 0 \end{vmatrix} = \bar{A}_1^2$, and setting $\bar{P}_1 = Y^{-1}P_1Y$, $\bar{P}_2 = XP_2X^{-1}$, we have from

$$(16) \quad \bar{A}_1^2 \bar{P}_1 = \bar{P}_2 \bar{A}_1^2$$

that \bar{P}_2 has form

$$\bar{P}_2 = \begin{vmatrix} P_{11} & P_{12} \\ 0 & P_{22} \end{vmatrix}.$$

We take $P_3 = P_4 = \dots = P_{p-1} = 0$, assume that the rank of A_3^2 is less than h_2 , and seek \bar{P}_1, \bar{P}_2 such that (16) is satisfied, and

$$\bar{P}_2 \bar{A}_3^2 = 0$$

where $\bar{A}_3^2 = XA_3^2$. Under our assumption that the rank of $A_3^2 < h_2$, we can find a vector $x = (0, 0, \dots, 0, x_i, x_{i+1}, \dots, x_{h_2}), x_i \neq 0$ which is annihilated by \bar{A}_3^2 . Choose \bar{P}_2 to have the vector x in its i^{th} row, and zeros in the other rows. Then $\text{trace } \bar{P}_2 = x_i$, so that \bar{P}_2 has a characteristic root $\neq 0$. Further, for this choice of \bar{P}_2 , we can always find \bar{P}_1 so that (16) is satisfied. Thus our assumption that A_3^2 has rank less than h_2 would lead to a matrix P which commutes with \mathfrak{M} and has x_i and 0 for characteristic roots, which gives a contradiction.

For the case that not all \mathfrak{F}_x belonging to the block \mathfrak{B} , appear in the indecomposable representation study of the relations (14) show that the \mathfrak{F}_x which do appear have multiplicity 1 and form a sequence

$$\mathfrak{F}_\alpha, \mathfrak{F}_{\alpha+1}, \dots, \mathfrak{F}_{\alpha+r}.$$

3) $l(\mathfrak{M}) = 3$. Here again the modular irreducible constituents of \mathfrak{M} must belong to a block of lowest kind. We have also here that $\mathfrak{M}^2\mathfrak{B} \neq 0$. Then there is a Cartan basis element $b_{\kappa\kappa} \in \mathfrak{M}^2$ such that $b_{\kappa\kappa}\mathfrak{B} \neq 0$, suppose that $b_{\kappa\kappa}x \neq 0$, $x \in \mathfrak{B}$. Since $b_{\kappa\kappa} = b_{\kappa, \kappa-1}b_{\kappa-1, \kappa}$ we have that $b_{\kappa-1, \kappa}x \neq 0$, similarly $b_{\kappa+1, \kappa}x \neq 0$, and from $b_{\kappa\kappa} = b_{\kappa\kappa}e_{\kappa}(11)$ we have also that $e_{\kappa}(11)x \neq 0$. Let, as above, \mathfrak{I}_{κ} denote the indecomposable left ideal generated by the elements (10). Then it may be shown that \mathfrak{I}_{κ} contains no left annihilators of x other than 0, so that

$$a \rightarrow ax, \quad a \in \mathfrak{I}_{\kappa}$$

is an isomorphic mapping of \mathfrak{I}_{κ} upon the invariant subspace $\mathfrak{B}_1 = \mathfrak{I}_{\kappa}x$ of \mathfrak{B} . It follows that \mathfrak{M} contains a constituent equivalent to the indecomposable \mathfrak{U}_{κ} of \mathfrak{R} which corresponds to \mathfrak{I}_{κ} . This implies that the indecomposable representation \mathfrak{M} is equal to \mathfrak{U}_{κ} .¹³

We gather these results in

THEOREM I. *An indecomposable representation \mathfrak{M} of \mathfrak{R}_m has Loewy length $l(\mathfrak{M}) \leq 3$. If $l(\mathfrak{M}) = 1$, \mathfrak{M} is equivalent to a modular irreducible part of \mathfrak{R}_m . For $l(\mathfrak{M}) = 2$, \mathfrak{M} has the form (12) with $h_i = 1$ for a sequence of consecutive values of i , and the remaining $h_i = 0$, or \mathfrak{M} is of the similar form but with the positions of the odd and the even constituents reversed. When $l(\mathfrak{M}) = 3$, \mathfrak{M} is equivalent to an indecomposable part of the regular representation of \mathfrak{R}_m .*

6. Nakayama's results

We have seen above that the number of blocks of lowest kind is P_l , and that the number of blocks of highest kind is $P_m - pP_l$. Nakayama (8 II) has shown how to relate blocks and partitions more precisely. In the present section we state his results and give a slight extension of them.

Associated with the partition $(\lambda): \lambda_1 + \dots + \lambda_k = m$, $(\lambda_1 \geq \dots \geq \lambda_k > 0)$ of m is a diagram T consisting of k rows of fields, λ_i fields in the i^{th} row, the j^{th} elements of the rows being arranged in a column. The field in the i^{th} row and j^{th} column will be denoted by (i, j) . By the (i, j) -hook $H = H(i, j)$ of T we mean the set of fields (i, v) with $v \geq j$ together with the fields (v, j) with $v > i$. We call the number, h , of fields in H its length. If just r rows of T contain elements of H we call r the height of H . By $T - H$ we mean the diagram T' obtained from T by deleting the fields of H and then moving each field (i', j') with $i' > i, j' > j$ one row up and one column to the left.

We next divide the partitions (λ) of m into classes according to the following rules. If the diagram, T , of (λ) has no hook of length p , then (λ) is put in a class by itself, called a class of highest kind. If T has a hook, H , of length p and height r , we denote (λ) by $\lambda^r(\mu)$ where (μ) is the partition whose diagram is $T - H$. For $r = 1, \dots, p$ there is exactly one partition $\lambda^r(\mu)$ for each partition (μ) of $l (= m - p)$. The p partitions $\lambda^r(\mu)$ defined by (μ) are put into a class which we call a class of lowest kind. We can now state Nakayama's results.

¹³ Cf. Remark 3, Nakayama-Nesbitt, [7].

THEOREM II. Let $\mathfrak{A}(\lambda)$ denote the ordinary irreducible representation of \mathfrak{S}_m defined by (λ) and let $\mathfrak{F}(\lambda)$ be the modular representation induced by $\mathfrak{A}(\lambda)$. If (λ) belongs to a class of highest kind, then $\mathfrak{F}(\lambda)$ is irreducible and constitutes a block of highest kind. The p representations $\mathfrak{A}(\lambda^r(\mu))$ $r = 1, \dots, p$, are the ordinary irreducible representations belonging to a block of lowest kind, which we accordingly denote by $\mathfrak{B}(\mu)$. We can enumerate the modular irreducible representations $\mathfrak{F}_1(\mu), \dots, \mathfrak{F}_{p-1}(\mu)$ belonging to $\mathfrak{B}(\mu)$ so that

$$\begin{aligned}\mathfrak{F}(\lambda^1(\mu)) &\leftrightarrow \mathfrak{F}_1(\mu), \mathfrak{F}(\lambda^2(\mu)) \leftrightarrow \mathfrak{F}_1(\mu) + \mathfrak{F}_2(\mu), \dots, \\ \mathfrak{F}(\lambda^{p-1}(\mu)) &\leftrightarrow \mathfrak{F}_{p-2}(\mu) + \mathfrak{F}_{p-1}(\mu), \mathfrak{F}(\lambda^p(\mu)) \leftrightarrow \mathfrak{F}_{p-1}(\mu)\end{aligned}$$

(\leftrightarrow denotes "has same irreducible constituents as" not "is equivalent to").

Let $(\mu) = (\mu_1, \dots, \mu_k)$ (where $\mu_1 \geq \mu_2 \geq \dots \geq \mu_k > 0 = \mu_{k+1} = \dots$) be a partition of l . If $\mu_i > \mu_{i+1}$ we denote by $(\mu | i)$ the partition $(\mu_1, \dots, \mu_i - 1, \dots, \mu_k)$ of $l - 1$. In the course of proving the above main theorem Nakayama showed that

$$(17) \quad (\lambda^j(\mu) | i) = \lambda^j(\mu | i)$$

except when (a) $i = 1$ and $\mu_j = \mu_{j+1}$ or (b) $i = j - 1$ and $\mu_j = \mu_{j-1}$. (In other words successive removal of hooks is almost a commutative process). This fact enables us to prove the following corollary¹⁴ to the main theorem. If we consider a representation \mathfrak{F} of \mathfrak{S}_m only for permutations omitting the letter m we get a representation of \mathfrak{S}_{m-1} which we denote by \mathfrak{F}^* .

COROLLARY I.

$$(18) \quad \mathfrak{F}_j(\mu)^* \leftrightarrow \sum_i \mathfrak{F}_j(\mu | i) + \delta_{\mu_j \mu_{j+1}} \mathfrak{F}(\mu_j + p - j, \mu_1 + 1, \dots, \mu_{j-1} + 1, \mu_{j+1}, \dots, \mu_k)$$

The sum is over all i for which $\mu_i > \mu_{i+1}$.

PROOF. It is well known¹⁵ that $\mathfrak{A}(\lambda)^* \leftrightarrow \sum_i \mathfrak{A}(\lambda | i)$ where the sum extends over all i for which $\lambda_i > \lambda_{i+1}$. Applying Theorem II and equation (17) to the induced modular representation $\mathfrak{F}(\lambda)^*$ for $(\lambda) = \lambda^p(\mu)$ we get

$$(19) \quad \begin{aligned}\mathfrak{F}(\lambda^p(\mu))^* &\leftrightarrow \sum_i \mathfrak{F}_{p-1}(\mu | i) + \mathfrak{F}_p(\mu | i) \\ &\quad + \delta_{\mu_p \mu_{p+1}} \mathfrak{F}(\lambda^p(\mu) | 1) + \delta_{\mu_{p-1} \mu_p} \mathfrak{F}(\lambda^p(\mu) | p - 1), \quad p = 1, \dots, p.\end{aligned}$$

(The sum is over all i for which $\mu_i > \mu_{i+1}$.) We have also from the main theorem that

$$(20) \quad \mathfrak{F}^j(\mu) \leftrightarrow \mathfrak{F}(\lambda^j(\mu)) - \mathfrak{F}(\lambda^{j-1}(\mu)) \dots (-1)^{j+1} \mathfrak{F}(\lambda^1(\mu))$$

The corollary now follows at once by starring both sides of (20), substituting from (19) in each term on the right hand side and simplifying.

¹⁴ Aside from this corollary everything in the present section is a restatement of Nakayama's results.

¹⁵ [13] p. 215.

Observe that the terms on the right hand side of (18) all come from different blocks and so we can strengthen the corollary by the additional statement that $\mathfrak{F}_j(\mu)^*$ is a completely reducible representation of \mathfrak{S}_{m-1} .

7. Definitions and notation

In the next section we will consider the decomposition of representations of \mathfrak{S}_m when considered as representations of various subgroups \mathfrak{S}_{m-v} . To facilitate this we introduce the following calculus of partitions.

Let (λ) be a partition of m (as in §6 above) and let $(i) = (i_1, \dots, i_v)$ be a sequence consisting of μ_1 1's, μ_2 2's, \dots , μ_k k 's. By $(\lambda | i) = (\lambda | i_1 \dots i_v)$ we mean the partition¹⁶ $(\lambda_1 - \mu_1, \dots, \lambda_k - \mu_k)$ of $m - v$. The sequence i_1, \dots, i_v is said to be (λ) -proper if for each ν from 1 to v the partition $(\lambda | i_1 \dots i_\nu)$ of $m - \nu$ has non-negative terms in non-increasing order. Otherwise (i) is called (λ) -improper.

In this and the following sections we shall understand by $\mathfrak{A}(\lambda): s \rightarrow A(\lambda \chi s)$ Young's rational semi-normal form¹⁷ of the irreducible representations of \mathfrak{S}_m defined by (λ) .

If the sequence (i) is (λ) -proper we define $\mathfrak{A}(\lambda | i)$ to be the corresponding representation $s \rightarrow A(\lambda | i \chi s)$ of \mathfrak{S}_{m-v} . If (i) is (λ) -improper we shall mean by $\mathfrak{A}(\lambda | i)$ a zero rowed square matrix. An important property of the semi-normal form is that $\mathfrak{A}(\lambda)$ when considered as a representation of \mathfrak{S}_{m-v} (the symmetric group on the first $m - v$ letters of \mathfrak{S}_m) is already in completely reduced form with the $\mathfrak{A}(\lambda | i)$ as diagonal constituents. In other words each (λ) -proper sequence (i) of length v contributes an irreducible constituent. If (i) and (i') are two (λ) -proper sequences of length v , then $\mathfrak{A}(\lambda | i)$ appears above $\mathfrak{A}(\lambda | i')$ if the first nonvanishing difference $i_1 - i'_1, \dots, i_v - i'_v$ is positive.

We recall that the rows of $\mathfrak{A}(\lambda)$ are in 1-1 correspondence with the regular diagrams belonging to (λ) . The representation $\mathfrak{A}(\lambda | i)$ occupies the (consecutive) rows whose corresponding diagrams have m in the i_1^{th} row, $m - 1$ in the i_2^{th} row, \dots , $m - v + 1$ in the i_v^{th} row. By $A(\lambda | i | j \chi s)$ we mean the rectangular submatrix of $A(\lambda \chi s)$ whose rows are those of $\mathfrak{A}(\lambda | i)$ and whose columns are those of $\mathfrak{A}(\lambda | j)$. By the v^{th} refinement of $\mathfrak{A}(\lambda)$ we mean that each $A(\lambda \chi s)$ is to be considered as a matrix whose elements are the submatrices $A(\lambda | i | j \chi s)$.

8. The representations of \mathfrak{S}_p

For $m = p$ every $A(\lambda \chi s)$ is p -integral. There is just one block $\mathfrak{B} = \mathfrak{B}(0)$ of lowest kind. From Theorem II it follows that $\mathfrak{A}(\lambda)$ belongs to \mathfrak{B} if and only if the diagram of (λ) is a hook, i.e. one of the partitions $(p - \rho, 1^\rho)$, $\rho = 0, \dots, p - 1$.

¹⁶ Note that $\lambda_i \geq \mu_i$ is not required since for some purposes partitions with negative summands are useful.

¹⁷ [14] pp. 217-88 or [12].

Let $(\lambda) = (p - \rho, 1^\rho)$ where $0 < \rho < p - 1$. There are just four (λ) -proper sequences of length 2. These are (in order) $(i^1) = (\rho + 1, \rho)$, $(i^2) = (\rho + 1, 1)$, $(i^3) = (1, \rho + 1)$, $(i^4) = (1, 1)$. Note that the diagram of $(\lambda | i^r)$ is a hook of length $p - 2$. We denote by (λ^ρ) the partition $(p - \rho - 1, 1^{\rho-1})$ of $p - 2$, and by g^ρ the degree of the representation $\mathfrak{A}(\lambda^\rho)$ of \mathfrak{S}_{p-2} . In this notation we have $\mathfrak{A}(\lambda | i^1) = \mathfrak{A}(\lambda^{\rho-1})$, $\mathfrak{A}(\lambda | i^2) = \mathfrak{A}(\lambda | i^3) = \mathfrak{A}(\lambda^\rho)$, and $\mathfrak{A}(\lambda | i^4) = \mathfrak{A}(\lambda^{\rho+1})$.

For elements of \mathfrak{S}_{p-2} the non diagonal parts of the second refinement of $\mathfrak{A}(\lambda)$ all vanish and along the main diagonal we have $\mathfrak{A}(\lambda^{\rho-1})$, $\mathfrak{A}(\lambda^\rho)$, $\mathfrak{A}(\lambda^\rho)$, $\mathfrak{A}(\lambda^{\rho+1})$ in the order named. Let t_r denote the transposition $(r - 1, r)$. Then $A(\lambda | i^j | i^k \chi t_{p-1}) = 0$ for $j \neq k$ unless $(j, k) = (1, 2), (2, 1), (3, 4)$ or $(4, 3)$; and $A(\lambda | i^j | i^k \chi t_p) = 0$ for $j \neq k$ unless $(j, k) = (2, 3)$ or $(3, 2)$, the non zero parts being scalar; say $A(\lambda | i^j | i^k \chi t_p) = \alpha_{jk} E_{jk}$ where E_{jk} is the identity matrix of suitable degree. For the α_{jk} we have $\alpha_{11} = -1$, $\alpha_{22} = -1/(p - 1)$, $\alpha_{23} = (p^2 - 2p)/(p - 1)^2$, $\alpha_{32} = 1$, $\alpha_{33} = 1/(p - 1)$, $\alpha_{44} = 1$.

Now consider the induced modular representation $\mathfrak{F}(\lambda)$. Since $\mathfrak{A}(\lambda)$ is p -integral we are justified in keeping the same refinement notation, i.e. we write $F(\lambda \chi s) = || F(\lambda | i^j | i^k \chi s) ||$ where each submatrix of $F(\lambda \chi s)$ is obtained by considering the corresponding submatrix of $A(\lambda \chi s) \bmod p$. For $s \in \mathfrak{S}_{p-1}$ we have

$$F(\lambda \chi s) = \left\| \begin{array}{cc|cc} F(\lambda | \rho + 1 \chi s) & 0 & 0 & \\ & 0 & 0 & \\ 0 & 0 & & F(\lambda | 1 \chi s) \\ 0 & 0 & & \end{array} \right\|$$

where $F(\lambda | \rho + 1 \chi s)$ occupies the first two sets of rows and columns and $F(\lambda | 1 \chi s)$ the last two sets. $F(\lambda | \rho + 1)$ is (as the notation implies) $A(\lambda | \rho + 1)$ taken mod p .

Furthermore, we have

$$F(\lambda \chi t_p) = \left\| \begin{array}{cccc} -E_{11} & 0 & 0 & 0 \\ 0 & E_{22} & 0 & 0 \\ 0 & E_{32} & -E_{33} & 0 \\ 0 & 0 & 0 & E_{44} \end{array} \right\|.$$

Since \mathfrak{S}_p is generated by \mathfrak{S}_{p-1} and t_p , the forms of these matrices show just how to write down the irreducible constituents $(\mathfrak{F}_\rho(0), \mathfrak{F}_{\rho+1}(0))$ of $\mathfrak{F}(\lambda)$. (For $\rho = 0$ and $\rho = p - 1$ the situation differs only in that certain indicated parts of the refinement do not appear, and of course there are no representations $\mathfrak{F}_\rho(-1)$ and $\mathfrak{F}_\rho(p)$). Summarizing we see that $\mathfrak{F}_\rho(0 \chi s) = \mathfrak{F}(p - \rho, 1^{\rho-1} \chi s)$ if $s \in \mathfrak{S}_{p-1}$ and $F_\rho(0 \chi t_p) = \left\| \begin{array}{cc} -E_{11} & 0 \\ 0 & E_{22} \end{array} \right\|$ where E_{11} is of degree $g^{\rho-1}$ and E_{22} of degree g^ρ .

This completes the determination of the irreducible representations of \mathfrak{S}_p .

But we can get still more information from the above process. For by Theorem I it follows that the lower left hand corner (last two sets of rows first two sets of columns) of $\mathfrak{F}(\lambda)$ must be a numerical multiple of $\mathfrak{S}_p^{\rho+1}(0)$. We may, and do, suppose that the Cartan basis for \mathfrak{R}_p was so chosen that this numerical multiple (which is obviously not zero) is unity. Then we have $H_p^{\rho+1}(0\check{s}) = 0$ if $s \in \mathfrak{S}_{p-1}$ and $H_p^{\rho+1}(0\check{t}_p) = \begin{vmatrix} 0 & E_{32} \\ 0 & 0 \end{vmatrix}$ (E_{32} of degree g^ρ).

To calculate $\mathfrak{S}_{p+1}^\rho(0)$ we replace the above used form of Young's semi-normal representation by the one generated by the transposes of the matrices $A(\lambda\check{t}_r)$ $r = 2, \dots, p$; and we get $H_{p+1}^\rho(0\check{t}_p) = \begin{vmatrix} 0 & 0 \\ E_{23} & 0 \end{vmatrix}$ (i.e. the transpose of $H_p^{\rho+1}(0\check{t}_p)$), and $H_{p+1}^\rho(0\check{s}) = 0$ if $s \in \mathfrak{S}_{p-1}$.

The final step in our program of determining all representations of \mathfrak{S}_p is the calculation for the elementary modules $\mathfrak{S}_p^\rho(0)$ of unmixed type. This calculation is based upon the form (formula (11) above) of $\mathfrak{U}_p(0)$ and the following two facts: (1) $t_r^2 = 1$ and (2) t_p commutes with every element of \mathfrak{S}_{p-2} . It follows at once from $t_r^2 = 1$, $H_p^{\rho+1}(0\check{t}_r) = H_{p+1}^\rho(0\check{t}_r) = 0$ that $H_p^\rho(0\check{t}_r) = 0$ for $r < p$. Since \mathfrak{S}_{p-1} is generated by t_2, \dots, t_{p-1} this shows that $H_p^\rho(0\check{s}) = 0$ if $s \in \mathfrak{S}_{p-1}$. It remains therefore only to calculate $H_p^\rho(0\check{t}_p)$.

Considered as a representation of \mathfrak{S}_{p-2} , $\mathfrak{U}_p(0)$ takes the form

$$\begin{vmatrix} \mathfrak{F}(\lambda^{\rho-1}) & & & & & \\ & \mathfrak{F}(\lambda^\rho) & & & & \\ & & \mathfrak{F}(\lambda^{\rho-2}) & & & \\ & & & \mathfrak{F}(\lambda^{\rho-1}) & & \\ & & & & \mathfrak{F}(\lambda^\rho) & \\ & & & & & \mathfrak{F}(\lambda^{\rho-1}) \\ & & & & & & \mathfrak{F}(\lambda^\rho) \end{vmatrix}$$

Since t_p commutes with \mathfrak{R}_{p-2} it now follows from Schur's Lemma that $H_p^\rho(0\check{t}_p) = \begin{vmatrix} \alpha_1 E_{11} & 0 \\ 0 & \alpha_2 E_{22} \end{vmatrix}$, (E_{11} and E_{22} as above). Now apply the condition $t_p^2 = 1$ and we get $\alpha_1 = -\alpha_2 = 1/2$.

9. Further specific results

The above methods can be applied without serious additional complications (save in notation) to obtain similar results for \mathfrak{S}_{p+1} and \mathfrak{S}_{p+2} . For \mathfrak{S}_{p+1} there is again just one block of lowest kind and the ordinary representations belonging to it are p -integral (i.e. when put in rational semi-normal form). For \mathfrak{S}_{p+2} there are two blocks of lowest kind, and although some of the ordinary semi-normal representations belonging to these blocks fail to be p -integral, they become so after a simple transformation which does not irreparably upset the refinement

into submatrices. But for \mathfrak{S}_{p+l} with $l > 2$ no such simple transformation into p -integral form has yet been found.

One might hope to get further information by starting with one of the known integral forms for the irreducible representations of \mathfrak{S}_m , rather than with Young's semi-normal form. The drawback to such a procedure is that these integral forms are not adapted for descent to the subgroups \mathfrak{S}_{m-v} of \mathfrak{S}_m . So when they are taken mod p their decomposition seems to be almost (if not fully) as difficult as that of the regular representation, and so there is no particular point in using them.

INSTITUTE FOR ADVANCED STUDY AND
UNIVERSITY OF MICHIGAN

BIBLIOGRAPHY

1. R. BRAUER AND C. NESBITT, *On the Modular Characters of Groups*, these Annals, (2), vol. 42, (1941), pp. 556-590.
2. R. BRAUER, *Investigations on Group Characters*, these Annals, (4), vol. 42, (1941), pp. 936-958.
3. R. BRAUER, *On the Cartan Invariants of Groups of Finite Order*, these Annals, (1), vol. 42, (1941), pp. 53-61.
4. R. BRAUER, *On Sets of Matrices with Coefficients in a Division Ring*, Trans. Amer. Math. Soc. (3), vol. 49, (1941), pp. 502-548.
5. R. BRAUER, *On the Representation of Groups of Finite Order*, Proceedings of the National Academy of Sciences, vol. 25, (1939), pp. 290-295.
6. C. NESBITT, *On the Regular Representations of Algebras*, these Annals, (3), vol. 39, (1938), pp. 634-658.
7. T. NAKAYAMA AND C. NESBITT, *Note on Symmetric Algebras*, these Annals, (3), vol. 39, (1938), pp. 659-668.
8. T. NAKAYAMA, *On Some Modular Properties of Irreducible Representations of a Symmetric Group I*, Japanese Journal of Math., vol. XVII, (1940), pp. 89-108; *II*, Japanese Journal of Math., vol. XVII, (1941), 411-423.
9. T. NAKAYAMA, *Some Studies on Regular Representations and Modular Representations*, these Annals, (2), vol. 39, (1938), pp. 361-369.
10. W. M. SCOTT, *On Matrix Algebras Over an Algebraically Closed Field*, these Annals (2), vol. 43, (1942), pp. 147-160.
11. R. BRAUER AND I. SCHUR, *Zum Irreduzibilitätsbegriff in der Gruppen linearer homogener Substitutionen*, Sitzungsber. Preuss. Akad., 1930, p. 209, §2.
12. R. M. THRALL, *On Young's Semi-normal Representation of the Symmetric Group*, Duke Journal, vol. 8, (1941), pp. 611-624.
13. H. WEYL, *The Classical Groups*, Princeton, 1939.
14. A. YOUNG, *Quantitative Substitutional Analysis, Part VI*, Proceedings of the London Mathematical Society (2), vol. 34, (1932), pp. 196-230.

ON THE DECOMPOSITION OF MODULAR TENSORS (I)

By R. M. THRALL

(Received December 11, 1941)

1. Introduction

This paper is a companion to the preceding one¹ [5], so we number formulas, theorems, etc., consecutively from those in that paper, and preserve the same notation unless a change is specifically indicated.

Let \mathfrak{V}_1 be a vector space over a field \mathfrak{f} of characteristic p . We are interested in the decomposition of the space \mathfrak{V}_m of all tensors of rank m , relative to the Kronecker m^{th} power representation Π_m of the group \mathfrak{G} of all non-singular linear transformations of \mathfrak{V}_1 into itself. In this paper we determine the structure of \mathfrak{V}_m subject to the two limitations: I. $m < 2p$; II. \mathfrak{f} has at least m elements. The first limitation is due to the incomplete state of the theory of modular representations of the symmetric group. The second limitation is less serious, although the decomposition is actually different if \mathfrak{f} has less than m elements. We hope to treat this case in a later paper.

The principal results about \mathfrak{V}_m are contained in Theorems III and VII, together with formula (38). The problem is attacked by exhibiting the enveloping algebra of Π_m as the commutator algebra of a certain permutation representation of the symmetric group of degree m . A main tool in this process is application of Remark I, below, which states that the order of the commutator algebra of a group of permutation matrices is independent of the underlying field, i.e. is even the same characteristic 0 as characteristic p .

2. The commutator algebra of a monomial group

An n -rowed matrix is called *monomial* if it has exactly one non-zero element in each row and column. A *permutation matrix* is a monomial matrix in which each of the n non-zero elements is 1. A *diagonal matrix* is a monomial matrix whose non-zero terms all lie on the main diagonal.

Let $\mathfrak{A}: s \rightarrow A(s) = ||a_{ij}(s)||$ be a monomial \mathfrak{f} -representation of degree n of a group \mathfrak{G} ; \mathfrak{f} being any field. (We call \mathfrak{A} monomial when each $A(s)$ is monomial.) The set \mathfrak{B} of all \mathfrak{f} -matrices B such that

$$(21) \quad A(s)B = BA(s) \quad \text{for all } s \text{ in } \mathfrak{G}$$

is called the *commutator algebra* of the representation \mathfrak{A} . We are interested in determining the nature and order of \mathfrak{B} .

We first treat the case in which \mathfrak{A} is a permutation representation. Suppose that

¹ Brackets refer to the bibliography at the end of the paper.

$$(22) \quad A(s) \begin{Bmatrix} x_1 \\ \vdots \\ x_n \end{Bmatrix} = \begin{Bmatrix} x_{1(s)} \\ \vdots \\ x_{n(s)} \end{Bmatrix}$$

then (21) in the form $A(s)BA(s)^{-1} = B$ is equivalent to the equations

$$(23) \quad b_{i(s)j(s)} = b_{ij} \quad i, j = 1, \dots, n, \text{ for all } s \text{ in } \mathfrak{G}.$$

We divide the index pairs (i, j) into systems of transitivity according to the equivalence relation:

$$(24) \quad (i, j) \sim (i', j') \leftrightarrow \text{there is an } s \text{ in } \mathfrak{G} \text{ for which } i' = i(s); j' = j(s).$$

It is clear from (23) that B is a commutator of \mathfrak{A} if and only if

$$(25) \quad b_{ij} = b_{i'j'} \text{ whenever } (i, j) \sim (i', j').$$

LEMMA I. *The order of the commutator algebra of a permutation representation (of a finite group) is equal to the number of systems of transitivity in the Kronecker square of the representation.*

PROOF. We can consider the n^2 numbers $x_i y_j$ as coordinates of the general vector in the space on which $\mathfrak{A} \times \mathfrak{A}$ operates. Then it follows from (22) that $A(s) \times A(s)$ sends the column matrix $\|x_i y_j\|$ into the column matrix $\|x_{i(s)} y_{j(s)}\|$. Thus the systems of transitivity under $\mathfrak{A} \times \mathfrak{A}$ are just the same as the systems of transitivity of the index pairs (i, j) defined by (24) above. With a system of transitivity we associate the matrix B having $b_{ij} = 1$ if (i, j) is in the given system and $b_{ij} = 0$ otherwise. The matrices thus constructed are clearly linearly independent; and it follows from (25) that they constitute a basis for \mathfrak{B} .

A monomial matrix A can be represented uniquely as the product DP of a diagonal matrix and a permutation matrix. If $A' = D'P'$ is a second monomial matrix, then the product $A'' = AA' = D''P''$ has $D'' = DPD'P^{-1}$ and $P'' = PP'$. Returning now to the arbitrary monomial representation \mathfrak{A} we write $A(s) = D(s)P(s)$; $D(s) = \|\delta_{ij} d_i(s)\|$. We have just proved that $P(st) = P(s)P(t)$ and so $s \rightarrow P(s)$ is a permutation representation \mathfrak{P} of \mathfrak{G} which we call the *permutation representation belonging to* \mathfrak{A} .

The equations (21) are now equivalent to

$$(26) \quad d_i(s)b_{i(s)j(s)}/d_j(s) = b_{ij} \quad \text{for all } s \text{ in } \mathfrak{G}, i, j = 1, \dots, m.$$

We divide the index pairs (i, j) into systems of transitivity according to \mathfrak{P} . Consider the subgroup $\mathfrak{S} = \mathfrak{S}(i, j)$ containing all s for which $i = i(s), j = j(s)$. We say that the system of transitivity containing (i, j) is *singular* if for some s in \mathfrak{S} , $d_i(s) \neq d_j(s)$. [A simple computation shows that being singular is a property of the system of transitivity, independent of the representative (i, j) used to test for singularity.] If (i, j) belongs to a singular system we have $b_{ij} = 0$ by (26) and then by transitivity $b_{i'j'} = 0$ for every $(i', j') \sim (i, j)$. However, if (i, j) belongs to a non-singular system then there is a commutator B

with $b_{ij} = 1$, and $b_{i'j'} \neq 0$ if and only if $(i', j') \sim (i, j)$. The equations (26) will never lead to relations connecting b_{ij} and $b_{i'j'}$ unless $(i, j) \sim (i', j')$, so if there is any solution with $b_{ij} = 1$, there is one with $b_{ij} = 1$ and $b_{i'j'} = 0$ if (i', j') is not in the same system as (i, j) . Since the $A(s)$ form a group, any equation in (26) connecting $b_{i(s)j(s)}$ and $b_{i(t)j(t)}$ is implied by those connecting b_{ij} to $b_{i'j'}$ with $(i', j') = (i(s), j(s)), (i(t), j(t)), (i(ts^{-1}), j(ts^{-1}))$. Hence our problem is reduced to showing that the equations (26) with b_{ij} on the right-hand side, are soluble with $b_{ij} = 1$. We know that for s in \mathfrak{S} they are consistent. If $(i(s), j(s)) = (i(t), j(t))$ then $u = ts^{-1} \in \mathfrak{S}$ and now calculating $d_i(t), d_j(t)$ from the equation $A(t) = A(u)A(s)$, we get $d_i(t)/d_j(t) = d_i(s)/d_j(s)$ which establishes the consistency of the equations. We have now proved

LEMMA II. *The order of the commutator algebra of a group of monomial matrices is equal to the number of non-singular transitive systems of index pairs.*

Let \mathfrak{f} and \mathfrak{K} be any two fields. A group of permutation matrices can be regarded as lying in either \mathfrak{f} or \mathfrak{K} , since 1 is an element of any field. Lemma I states that the order of the \mathfrak{f} -commutator algebra is equal to the order of the \mathfrak{K} -commutator algebra. We shall apply this fact below to the case where one field is of characteristic 0 and the other of characteristic p . For future reference we restate this (weaker) form of Lemma I as

REMARK I. *The order of the commutator algebra of a group of permutation matrices is independent of the field of coefficients.*

The analogue to Remark I for monomial groups is not true. For consider the group of order p generated by $A(s) = \begin{vmatrix} w & 0 \\ 0 & 1 \end{vmatrix}$ where w is a p^{th} root of unity (in a field of characteristic 0). Since $w \equiv 1 \pmod{p}$, the modular image of this group is generated by $A(s) = \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix}$. The non-modular commutator algebra has order 2 and the modular one order 1.

3. Vanishing forms

Suppose that \mathfrak{f} is the Galois field with q elements. Let Y_1, Y_2, \dots be indeterminants over \mathfrak{f} , and let y_1, y_2, \dots be variables whose domain is \mathfrak{f} . Consider the ideal in $\mathfrak{f}[Y_1, Y_2, \dots]$ generated by $Y_i^q - Y_i, i = 1, 2, \dots$. Modulo this ideal every \mathfrak{f} -polynomial $F(Y_1, Y_2, \dots)$ is congruent to a unique \mathfrak{f} -polynomial $F^*(Y_1, Y_2, \dots)$ of degree less than q in each Y_i , and $F(y_1, y_2, \dots) = 0$ is equivalent to $F^*(Y_1, Y_2, \dots) = 0$. In particular, a \mathfrak{f} -polynomial $F(Y_1)$ of degree less than q cannot vanish over \mathfrak{f} (i.e. $F(y_1) = 0$) unless every coefficient is zero.

LEMMA III. *There is no non-zero \mathfrak{f} -form $P(Y_{ij})$ of degree $m \leq q$ in the n^2 indeterminants Y_{ij} , such that $P(y_{ij}) \det ||y_{ij}|| = 0$ where the y_{ij} are n^2 variables whose domain is \mathfrak{f} .*

PROOF. If $m < q - 1$, then $F(Y_{ij}) = P(Y_{ij}) \det ||Y_{ij}||$ is of degree less than q in each Y_{ij} , so $P(Y_{ij}) \neq 0$ implies $F(Y_{ij}) = F^*(Y_{ij}) \neq 0$ and therefore $F(y_{ij}) \neq 0$. This leaves the two cases $m = q - 1, m = q$. The proofs for

these are similar, so we treat here only the harder case, $m = q$. Write $P(Y_{ij})$ as a polynomial in Y_{11} with coefficients $P_i = P_i(Y_{ij})$ that are polynomials in $\mathfrak{f}[Y_{12}, \dots, Y_{nn}]$, i.e.

$$P(Y_{ij}) = P_0 Y_{11}^m + P_1 Y_{11}^{m-1} + \dots + P_m.$$

We also write

$$\det || Y_{ij} || = Q_0 Y_{11} + Q_1.$$

Then

$$F(Y_{ij}) = Q_0 P_0 Y_{11}^{m+1} + (Q_0 P_1 + Q_1 P_0) Y_{11}^m + \dots + Q_1 P_m.$$

If we replace Y^{m+1} by Y^2 and Y^m by Y , $F(Y_{ij})$ is replaced by the congruent polynomial (no longer a form)

$$\begin{aligned} F'(Y_{ij}) &= (Q_0 P_2 + Q_1 P_1) Y_{11}^{m-1} + \dots + (Q_0 P_{m-2} + Q_1 P_{m-3}) Y_{11}^3 \\ &+ (Q_0 P_0 + Q_0 P_{m-1} + Q_1 P_{m-2}) Y_{11}^2 + (Q_0 P_1 + Q_1 P_0 + Q_1 P_{m-1}) Y_{11} + Q_1 P_m. \end{aligned}$$

Since $F'(Y_{ij})$ is of degree $q - 1$ (or less) in Y_{11} we cannot have $F'(y_{ij}) = 0$ unless the coefficient of each power of Y_{11} becomes zero when Y_{ij} is replaced by y_{ij} . For $i > 2$ the coefficient of Y_{11}^i is of degree less than q in each indeterminate and so vanishing for y_{ij} is the same as vanishing for Y_{ij} ; i.e.

$$(27) \quad Q_0 P_i = -Q_1 P_{i-1} \quad i = 2, \dots, m-2.$$

The lemma is trivial if $n = 1$. For $n > 1$, Q_0 and Q_1 are relatively prime forms of degree > 1 . (27) for $i = 2$ requires $P_1 = P_2 = 0$, for otherwise P_1 , of degree 1, would be divisible by Q_0 , of degree > 1 . Hence, $P_1 = P_2 = \dots = P_{m-2} = 0$. We can apply the same process to each Y_{ij} and conclude that $F(y_{ij}) = 0$ implies that $P(Y_{ij})$ has no terms of degree different from m , 1, 0 in each Y_{ij} . But then $F^*(Y_{ij}) = P^*(Y_{ij}) \det || Y_{ij} ||$ and $P^*(Y_{ij})$ is clearly not zero unless $P(Y_{ij}) = 0$; this completes the proof for $m = q$. Observe that the form $P(Y) = Y_{11}^q Y_{12} - Y_{11} Y_{12}^q$ shows that the lemma is false for $m > q$.

We use Lemma III as a modular substitute for what H. Weyl [6, p. 4] calls the "principle of the irrelevance of algebraic inequalities."

4. General program

Henceforth \mathfrak{f} shall be a field of characteristic p . We are interested in the decomposition of the Kronecker m^{th} power representation $\Pi_m: A \rightarrow \Pi_m(A) = A \times \dots \times A$ (m -factors) of the full linear group $\mathfrak{G} = \mathfrak{G}(n, \mathfrak{f})$ of n -rowed non-singular \mathfrak{f} -matrices. We propose to effect this decomposition by exhibiting the enveloping algebra \mathfrak{A}_m of Π_m as the commutator algebra of a certain permutation representation of the symmetric group \mathfrak{S}_m of degree m .

The order of \mathfrak{A}_m is the number of linearly independent monomials of degree m in $N = n^2$ variables a_{ij} which range freely over \mathfrak{f} save for the restriction $\det || a_{ij} || \neq 0$. Hence, by Lemma III we have

LEMMA IV. *If \mathfrak{f} has at least m elements, then the order of \mathfrak{A}_m is $\binom{N+m-1}{m}$; i.e. is equal to the number of linearly independent monomials of degree m in N indeterminants.*

Let $x(i_1, \dots, i_m)(i_1, \dots, i_m = 1, \dots, n)$ be the components of an arbitrary vector in the \mathfrak{f} -space \mathfrak{B}_m of all tensors of rank m ; i.e. \mathfrak{B}_m is the representation space [6, pp. 96–98] for the Kronecker m^{th} power representation of \mathfrak{G} . Let s be the permutation $1 \rightarrow 1', \dots, m \rightarrow m'$. We define s as a linear operator on \mathfrak{B}_m by the equation $x' = sx$ where $x'(i_1, \dots, i_m) = x(i_{1'}, \dots, i_{m'})$. Let $T_m(s)$ be the matrix which describes this mapping. It is evident that $\mathfrak{T}_m : s \rightarrow T_m(s)$ is a permutation representation of degree n^m of \mathfrak{S}_m .

We call a \mathfrak{f} -matrix of degree n^m *bisymmetric* if it commutes with every $T_m(s)$. Since the set \mathfrak{B}_m of all bisymmetric \mathfrak{f} -matrices is the commutator algebra of a permutation representation, its order will be the same as the order of the set \mathfrak{B}_m^0 of all bisymmetric matrices in a field of characteristic 0. This latter order [6, p. 130] is known to be $\binom{N+m-1}{m}$.

It is trivial to verify that $\Pi_m(A)$ is bisymmetric; i.e. $\mathfrak{A}_m \subseteq \mathfrak{B}_m$. Now apply Lemma IV and we see that

THEOREM III. *If \mathfrak{f} has at least m elements, then the enveloping algebra of the Kronecker m^{th} power representation of the full linear group of degree n is the set of all bisymmetric \mathfrak{f} -matrices of degree n^m .*

Still paralleling the non-modular theory, our next step is to determine the indecomposable constituents of \mathfrak{T}_m . Then we obtain the decomposed form of \mathfrak{A}_m , by starting with the commutator algebra of the decomposed form of \mathfrak{T}_m .

When this is all accomplished we shall know the structure of the decomposed form of Π_m ; i.e. we shall know the degrees of the irreducible constituents; the nature and multiplicities of the indecomposable constituents. But we shall still have no direct construction for the representations themselves, or much information about the characters of the representations. This is a general difficulty encountered when a representation of a group is studied by determining its enveloping algebra as a commutator algebra. The root of this difficulty is the lack of criteria for determining which elements of the enveloping algebra actually correspond to group elements. Attempts to remedy these deficiencies for the present theory are postponed to later papers. We also postpone any discussion of the case in which \mathfrak{f} has less than m elements.

5. The representations \mathfrak{T}_m

We may regard \mathfrak{T}_m as a permutation group whose elements $T_m(s)$ are written on the "letters" $x(i_1, \dots, i_m)$. Two letters $x(i_1, \dots, i_m)$ and $x(j_1, \dots, j_m)$ belong to the same system of transitivity of \mathfrak{T}_m if and only if the integers j_1, \dots, j_m are just the integers i_1, \dots, i_m in some arrangement. We now arrange the basis vectors of \mathfrak{B}_m so that the letters of the several systems of transitivity are brought together. In the language of representation theory,

this exhibits the representation \mathfrak{T}_m as the direct sum [6, pp. 19, 20] of its *transitive constituents*.

Let $(\lambda) = (\lambda_1, \dots, \lambda_k)$ denote the partition $m = \lambda_1 + \dots + \lambda_k$, $\lambda_1 \geq \dots \geq \lambda_k > 0$, and let $x(i_1, \dots, i_m)$ have the first λ_1 indices all 1, \dots , the last λ_k indices all k . The permutations s such that $sx(i_1, \dots, i_m) = x(i_1, \dots, i_m)$ constitute a subgroup of \mathfrak{S}_m isomorphic to $\mathfrak{S}(\lambda) = \mathfrak{S}_{\lambda_1} \times \dots \times \mathfrak{S}_{\lambda_k}$ (here \times denotes group-theoretic direct product). Let $S(\lambda)$ denote the sum of the elements of $\mathfrak{S}(\lambda)$, and suppose that $s_1 = 1, s_2, \dots$ are elements, one from each coset of $\mathfrak{S}(\lambda)$ in \mathfrak{S}_m . Then the elements $s_i S(\lambda)$ constitute a \mathfrak{f} -basis for the left ideal $\mathfrak{L}(\lambda)$ of \mathfrak{R}_m (the \mathfrak{f} group ring of \mathfrak{S}_m) generated by $S(\lambda)$. Left multiplication of $\mathfrak{L}(\lambda)$ by a permutation s merely permutes the basis elements $s_i S(\lambda)$, and so $\mathfrak{L}(\lambda)$ is representation space for a (transitive) permutation representation [3, p. 110] $\mathfrak{T}_{(\lambda)} : s \rightarrow T_{(\lambda)}(s)$ of \mathfrak{S}_m .

It is obvious that, as left \mathfrak{R}_m -space $\mathfrak{L}(\lambda)$ is isomorphic to the \mathfrak{f} -space made up of tensors whose only non-zero coordinates are $x(i_1, \dots, i_m)$ and its conjugates $sx(i_1, \dots, i_m)$. Hence $\mathfrak{T}_{(\lambda)}$ is equivalent to a constituent of \mathfrak{T}_m ; and, conversely, it is clear that every transitive constituent of \mathfrak{T}_m is equivalent to one of the $\mathfrak{T}_{(\lambda)}$. So to know the decomposition of \mathfrak{T}_m we need only know the decomposition of each $\mathfrak{T}_{(\lambda)}$; or equivalently, to write each $\mathfrak{L}(\lambda)$ as a direct sum of indecomposable left ideals of \mathfrak{R}_m .

We have $S(\lambda)^2 = n(\lambda)S(\lambda)$ where $n(\lambda) = \lambda_1! \dots \lambda_k!$. Hence if λ_1 (and therefore every λ_i) is less than p , the ideal $\mathfrak{L}(\lambda)$ has the idempotent generator $e(\lambda) = S(\lambda)/n(\lambda)$; and so $\mathfrak{L}(\lambda) = \mathfrak{R}_m S(\lambda)$ can be written as the direct sum of indecomposable left ideals which are direct summands of \mathfrak{R}_m (i.e. which themselves have idempotent generators). Stating this in the language of representation theory we have

THEOREM IV. *If $\lambda_1 < p$, $\mathfrak{T}_{(\lambda)}$ is a direct sum of indecomposable constituents of the regular representation of \mathfrak{S}_m .*

If $m = p$, Theorem IV covers all but one partition, the exception being $\lambda_1 = p$. But $\mathfrak{T}_{(p)}$ is the identity representation, and so we have established the following theorem for $m = p$:

THEOREM V. *For $m < 2p$ an indecomposable constituent of \mathfrak{T}_m is either an indecomposable constituent of the regular representation of \mathfrak{S}_m , or one of the irreducible representations² $\mathfrak{F}_1(\mu)$ of \mathfrak{S}_m .*

PROOF. There is nothing to prove for $m < p$. We proceed by an induction on m based upon the already verified case $m = p$. We suppose $p < m < 2p$ and that the theorem is already verified for $m - 1$. Considered only for elements of \mathfrak{S}_{m-1} , \mathfrak{T}_m is just \mathfrak{T}_{m-1} repeated n times. Hence, any indecomposable constituent of \mathfrak{T}_m must, when considered only for elements of \mathfrak{S}_{m-1} , split into indecomposable constituents of \mathfrak{T}_{m-1} .

Reference to Theorem I shows that Theorem V is false only if (I) \mathfrak{T}_m has

² See Theorem I [1], p. 9.

³ The notation $\mathfrak{F}_1(\mu)$ is explained in [5], §6.

$\mathfrak{F}_j(\mu)$, $j > 1$, as an indecomposable direct constituent, or (II) \mathfrak{T}_m has an indecomposable direct constituent of Loewy length 2. We now apply Corollary I to show that either I or II contradicts our induction hypothesis.

The application to I is immediate. For let (μ) be a partition of $l = m - p$. Then since $m > p$, there will be some i for which $\mu_i > \mu_{i+1}$; and so $\mathfrak{F}_j(\mu)$ in \mathfrak{T}_m would require $\mathfrak{F}_j(\mu | i)$ in \mathfrak{T}_{m-1} , contrary to our induction hypothesis.

Let \mathfrak{B} be an indecomposable representation of \mathfrak{S}_m of Loewy length 2, whose irreducible constituents are $\mathfrak{F}_j(\mu)$, $j = j_0, \dots, j_0 + r$, $r \geq 1$. Let \mathfrak{B}^* denote \mathfrak{B} considered only for elements of \mathfrak{S}_{m-1} . If \mathfrak{B} is an indecomposable direct constituent of \mathfrak{T}_m , then \mathfrak{B}^* is a direct constituent of \mathfrak{T}_{m-1} .

By Corollary I, the irreducible constituents of \mathfrak{B}^* will be either of highest kind or of the form $\mathfrak{F}_j(\mu | i)$ for j in the range $j_0, \dots, j_0 + r$. Since no $\mathfrak{F}_j(\mu)$ is repeated in \mathfrak{B} , it follows from Corollary I that no $\mathfrak{F}_j(\mu | i)$ can be repeated in any indecomposable constituent of \mathfrak{B}^* ; and so \mathfrak{B}^* can contain no indecomposable constituent of Loewy length 3. Since $r \geq 1$ and $m > p$, \mathfrak{B}^* must contain some $\mathfrak{F}_j(\mu | i)$ for $j > 1$. This $\mathfrak{F}_j(\mu | i)$ must lie in an indecomposable direct constituent of \mathfrak{B}^* of Loewy length 1 or 2. But then \mathfrak{B}^* cannot be a direct constituent of \mathfrak{T}_{m-1} , because of our induction hypothesis; hence \mathfrak{B} cannot be a direct constituent of \mathfrak{T}_m ; i.e. II is impossible. This completes the proof of Theorem V.

6. The commutator algebra of \mathfrak{T}_m

Instead of studying \mathfrak{T}_m itself, we investigate the more general case of any representation which is the sum of any number of the representations $\mathfrak{T}_{(\lambda)}$, repetitions permitted. Let \mathfrak{B} denote the decomposed form of any such representation. We group the constituents of \mathfrak{B} in such a way that

$$(28) \quad \mathfrak{B} = \left\| \begin{array}{c} \mathfrak{B}_1 \\ \mathfrak{B}_2 \\ \vdots \end{array} \right\|$$

where \mathfrak{B}_r consists of all the constituents of \mathfrak{B} that belong to the block \mathfrak{B}_r of \mathfrak{K}_m . The blocks \mathfrak{B}_r correspond to two-sided ideals that are minimal direct summands of \mathfrak{K}_m . Hence the commutator algebra \mathfrak{W} of \mathfrak{B} is the direct sum of the commutator algebras \mathfrak{W}_r of the \mathfrak{B}_r i.e.

$$(29) \quad \mathfrak{W} = \left\| \begin{array}{c} \mathfrak{W}_1 \\ \mathfrak{W}_2 \\ \vdots \end{array} \right\|$$

If \mathfrak{B}_r belongs to a block of lowest kind, it will just be an $\mathfrak{F}(\lambda)$ repeated, say, δ times. Then,⁴ since $\mathfrak{F}(\lambda)$ is a total matrix algebra, \mathfrak{W}_r is equivalent to the

⁴ Cf. [6], p. 92.

total \mathfrak{f} -matrix algebra of degree δ repeated $f(\lambda)$ times, where $f(\lambda)$ is the degree of $\mathfrak{F}(\lambda)$.

If \mathfrak{B}_r belongs to a block $\mathfrak{B}_r = \mathfrak{B}(\mu)$ of highest kind, the situation is somewhat more complicated. For simplicity in notation we drop all arguments (μ) from the letters denoting representations and matrices. According to Theorem V, \mathfrak{B}_r has the form:

$$(30) \quad \mathfrak{B}_r = \left\| \begin{array}{cccc} E_{\gamma_0} \times \mathfrak{F}_1 & & & \\ & E_{\gamma_1} \times \mathfrak{U}_1 & & \\ & & \ddots & \\ & & & E_{\gamma_{p-1}} \times \mathfrak{U}_{p-1} \end{array} \right\|$$

where E_r is the unit matrix of degree r and \times denotes Kronecker product. To obtain uniformity in notation we set $\mathfrak{F}_1 = \mathfrak{U}_0$. Suppose that $W_{ij}U_j(s) = U_i(s)W_{ij}$, for all s in \mathfrak{S}_m , and denote by A_{ij} any \mathfrak{f} -matrix of γ_i rows and γ_j columns, $i, j = 0, \dots, p-1$. Then

$$(31) \quad W_r = \left\| \begin{array}{cccc} A_{00} \times W_{00} & \cdots & A_{0p-1} \times W_{0p-1} \\ \cdots & & \\ A_{p-10} \times W_{p-10} & \cdots & A_{p-1p-1} \times W_{p-1p-1} \end{array} \right\|$$

is an element of \mathfrak{B}_r ; and, conversely, any element of \mathfrak{B}_r can be written as a linear \mathfrak{f} -combination of elements of the form (31).

To determine the number h_{ij} of linearly independent W_{ij} we use the Cartan matrix for the block \mathfrak{B} and apply the general theory of intertwining matrices.⁵ The result is that $h_{ij} = 0$ (and therefore $W_{ij} = 0$) unless j is $i-1$, i , or $i+1$. $h_{00} = 1$; $h_{ii} = 2$, $i = 1, \dots, p-1$; $h_{i,i+1} = h_{i+1,i} = 1$, $i = 0, \dots, p-1$. Since the A_{ij} 's are arbitrary this gives

$$(32) \quad \gamma_0^2 + 2\gamma_0\gamma_1 + 2\gamma_1^2 + 2\gamma_1\gamma_2 + \cdots + 2\gamma_{p-1}^2 \\ = (\gamma_0 + \gamma_1)^2 + (\gamma_1 + \gamma_2)^2 + \cdots + \gamma_{p-1}^2$$

for the order of \mathfrak{B}_r .

To determine the structure of \mathfrak{B}_r we must know the form of all the W_{ij} . We subdivide W_{ij} so that the rows of its submatrices are the same as the rows occupied by the irreducible constituents of \mathfrak{U}_i and the columns of the submatrices are the same as the rows occupied by the irreducible constituents of \mathfrak{U}_j . See formula (11) for the form of the \mathfrak{U}_i . We omit the details of the

⁵ See, for instance, Theorem 4 [4], p. 648.

computation. f_i denotes the degree of \mathfrak{F}_i , 0 stands for the zero matrix of proper size.

$$\begin{aligned}
 W_{00} &= \| aE_{f_1} \|, & W_{01} &= \| aE_{f_1} \quad 0 \quad 0 \|, \\
 W_{10} &= \left\| \begin{array}{c} 0 \\ 0 \\ aE_{f_1} \end{array} \right\|, & W_{ii} &= \left\| \begin{array}{cccc} aE_{f_1} & 0 & 0 & 0 \\ 0 & aE_{f_{i-1}} & 0 & 0 \\ 0 & 0 & aE_{f_{i+1}} & 0 \\ bE_{f_i} & 0 & 0 & aE_{f_i} \end{array} \right\|, \\
 W_{ii+1} &= \left\| \begin{array}{cccc} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ aE_{f_{i+1}} & 0 & 0 & 0 \\ 0 & aE_{f_i} & 0 & 0 \end{array} \right\|, & W_{i+1,i} &= \left\| \begin{array}{cccc} 0 & 0 & 0 & 0 \\ aE_{f_i} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & aE_{f_{i+1}} & 0 \end{array} \right\|
 \end{aligned}
 \tag{33}$$

The a 's appearing in different matrices W_{ij} are totally unrelated; for $i = 1$ and $i = p - 1$ the rows and columns of the W_{ij} that correspond formally to \mathfrak{F}_0 and \mathfrak{F}_p are to be deleted.

After a suitable shifting of rows and columns, \mathfrak{B} , takes the form

$$\left\| \begin{array}{c} E_{f_1} \times \mathfrak{B}^1 \\ \vdots \\ E_{f_{p-1}} \times \mathfrak{B}^{p-1} \end{array} \right\|
 \tag{34}$$

where \mathfrak{B}^i is the indecomposable matrix algebra given by

$$\mathfrak{B}^i = \left\| \begin{array}{cccc} \mathfrak{F}^i & & & \\ \mathfrak{F}^{i-1} & \mathfrak{F}^{i-1} & & \\ \mathfrak{F}^{i+1} & 0 & \mathfrak{F}^{i+1} & \\ \mathfrak{F}^i & \mathfrak{F}^{i-1} & \mathfrak{F}^{i+1} & \mathfrak{F}^i \end{array} \right\| \quad i = 1, \dots, p-1.
 \tag{35}$$

(For $i = p - 1$ the rows and columns of \mathfrak{F}^p are deleted.) Each \mathfrak{F}^i is a total \mathfrak{k} -matrix algebra of degree γ_i and the \mathfrak{F}_j^i are total \mathfrak{k} -matrix sets of degrees indicated by their positions in \mathfrak{B}^i .

It is of some interest to observe that if (I): $\gamma_i \neq 0$ implies $\gamma_r \neq 0$ for all $r < i$, holds for \mathfrak{B} , then the commutator algebra of \mathfrak{B} , is just the enveloping algebra of \mathfrak{B} .

7. Multiplicities of the constituents of \mathfrak{X}_m

We continue to regard $\mathfrak{X}_{(\lambda)}$, \mathfrak{X}_m , as permutation representations of \mathfrak{S}_m with the given \mathfrak{k} of characteristic p as underlying field. We shall denote by $\mathfrak{X}_{(\lambda)}^0$, \mathfrak{X}_m^0 the same permutation representations of \mathfrak{S}_m , but now regarded as made

up of matrices in a field \mathfrak{K} of characteristic 0. The structure of the representations $\mathfrak{T}_{(\lambda)}^0$, \mathfrak{T}_m^0 is well known.⁶ Denote by $\delta(\lambda)$ the multiplicity of $\mathfrak{A}(\lambda)$ in T_m^0 ; by $\delta(\lambda\check{\lambda}')$ the multiplicity of $\mathfrak{A}(\lambda)$ in $\mathfrak{T}_{(\lambda')}^0$.

If the diagram of (λ) has no hook of length p , i.e. if $\mathfrak{F}(\lambda)$ is of highest kind, then $\mathfrak{F}(\lambda)$ occurs $\delta(\lambda)$ times in \mathfrak{T}_m and $\delta(\lambda\check{\lambda}')$ times in $\mathfrak{T}_{(\lambda')}$. If the diagram of (λ) has a hook of length p , say $(\lambda) = \lambda^i(\mu)$ then we set $\delta(\lambda) = \delta_i(\mu)$ and $\delta(\lambda\check{\lambda}') = \delta_i(\mu\check{\lambda}')$. We let $\gamma_0(\mu), \gamma_1(\mu), \dots, \gamma_{p-1}(\mu)$ denote the multiplicities in \mathfrak{T}_m of $\mathfrak{F}_1(\mu), \mathfrak{U}_1(\mu), \dots, \mathfrak{U}_{p-1}(\mu)$ respectively; and let $\gamma_0(\mu\check{\lambda}'), \dots, \gamma_{p-1}(\mu\check{\lambda}')$ denote the corresponding multiplicities in $\mathfrak{T}_{(\lambda')}$. In formulas where they appear formally we set $\gamma_p(\mu) = \gamma_p(\mu\check{\lambda}) = 0$.

We get relations between the δ 's and the γ 's by counting the multiplicity of the modular irreducible constituents in two ways. Considering \mathfrak{T}_m as the modular representation induced by \mathfrak{T}_m^0 , it follows from the form of the decomposition matrix⁷ $D(\mu)$ for the block $\mathfrak{B}(\mu)$, that $\mathfrak{F}_i(\mu)$ occurs $\delta_i(\mu) + \delta_{i+1}(\mu)$ times in \mathfrak{T}_m . On the other hand, using the Cartan matrix $C(\mu)$ for the block $\mathfrak{B}(\mu)$ we see that $\mathfrak{F}_i(\mu)$ occurs $\gamma_{i-1}(\mu) + 2\gamma_i(\mu) + \gamma_{i+1}(\mu)$ times in \mathfrak{T}_m . Hence

$$(36) \quad \delta_i(\mu) + \delta_{i+1}(\mu) = \gamma_{i-1}(\mu) + 2\gamma_i(\mu) + \gamma_{i+1}(\mu), \quad i = 1, \dots, p-1.$$

The same reasoning applies to $\mathfrak{T}_{(\lambda)}$ giving

$$(37) \quad \begin{aligned} \delta_i(\mu\check{\lambda}) + \delta_{i+1}(\mu\check{\lambda}) &= \gamma_{i-1}(\mu\check{\lambda}) + 2\gamma_i(\mu\check{\lambda}) \\ &\quad + \gamma_{i+1}(\mu\check{\lambda}), \quad i = 1, \dots, p-1. \end{aligned}$$

We can now state the main theorem on multiplicities:

THEOREM VI. *If the diagram of (λ) has no hook of length p then $\mathfrak{F}(\lambda)$ occurs exactly as often in \mathfrak{T}_m as $\mathfrak{A}(\lambda)$ occurs in \mathfrak{T}_m^0 . For a block $\mathfrak{B}(\mu)$ of lowest kind we have:*

$$(38) \quad \begin{aligned} \gamma_{p-1}(\mu) &= \delta_p(\mu) \\ \gamma_{p-2}(\mu) &= \delta_{p-1}(\mu) - \delta_p(\mu) \\ &\dots\dots\dots \\ \gamma_{p-i}(\mu) &= \delta_{p-i+1}(\mu) - \delta_{p-i}(\mu) + \dots + (-1)^{i+1}\delta_p(\mu). \end{aligned}$$

Only the second part requires additional proof. If we add to equations (36) the single equation

$$(39) \quad \delta_1(\mu) = \gamma_0(\mu) + \gamma_1(\mu)$$

then a simple computation shows that (38) is the only solution of the augmented system. Hence to prove our theorem it is sufficient to establish (39). We do this by showing that

$$(40) \quad \delta_1(\mu\check{\lambda}) = \gamma_0(\mu\check{\lambda}) + \gamma_1(\mu\check{\lambda})$$

⁶ [3], Chapter IV, especially pp. 110, 128, 129; [6], Chapter VII, §§5, 6, 7.

⁷ [5], formulas (3) and (4).

for every partition (λ) of m , and then using the fact that \mathfrak{T}_m is a direct sum of constituents $\mathfrak{T}_{(\lambda)}$. Of course (40) in the presence of (37) leads to the following analogue of (38):

$$(41) \quad \gamma_{p-i}(\mu \check{\lambda}) = \delta_{p-i+1}(\mu \check{\lambda}) - \delta_{p-i}(\mu \check{\lambda}) + \cdots + (-1)^{i+1} \delta_p(\mu \check{\lambda}), \quad i = 1, \dots, p.$$

We arrange the partitions (λ) of m and the partitions (μ) of $l = m - p$ in dictionary order (i.e. (λ) precedes (λ') if the first non-vanishing difference $\lambda_i - \lambda'_i$ is positive). The verification of (40) is an induction argument, along the following lines: We suppose that (40) has been established for all $\mathfrak{T}_{(\lambda)}$ provided (μ) precedes a given (μ^0) . Then we exhibit one (λ^0) (actually $\lambda^1(\mu^0)$) such that $\mathfrak{F}_1(\mu^0)$ is the only constituent of $\mathfrak{B}(\mu^0)$ which appears in $\mathfrak{T}_{(\lambda^0)}$. Let $N(*)$ denote the order of the commutator algebra of any representation, $*$, of \mathfrak{S}_m . Then by Remark I we have

$$(42) \quad N(\mathfrak{B}) = N(\mathfrak{B}^0)$$

where \mathfrak{B} stands for any sum of $\mathfrak{T}_{(\lambda)}$ and \mathfrak{B}^0 is the sum of the corresponding $\mathfrak{T}_{(\lambda^0)}$. We establish (40) by solving for $\delta_1(\mu^0 \check{\lambda})$ in the equations obtained from (42) by setting $\mathfrak{B} = \mathfrak{T}_{(\lambda)}, \mathfrak{T}_{(\lambda^0)}, \mathfrak{T}_{(\lambda)} + \mathfrak{T}_{(\lambda^0)}$ in turn.

An important step in this process is the proof that a suitable (λ^0) exists. To accomplish this we analyze the character⁸ of $\mathfrak{T}_{(\lambda^0)}$, where (λ') is any partition of m , and obtain the following

LEMMA V. $\mathfrak{A}(\lambda')$ occurs exactly once as a constituent of $\mathfrak{T}_{(\lambda^0)}$, and $\mathfrak{A}(\lambda)$ cannot be a constituent of $\mathfrak{T}_{(\lambda^0)}$ if (λ') precedes (λ) ; i.e.

$$(43) \quad \delta(\lambda' \check{\lambda}) = 1, \quad \delta(\lambda \check{\lambda}) = 0 \quad \text{unless } (\lambda) \text{ precedes } (\lambda').$$

We observe that if (μ^0) precedes (μ) then $\lambda^1(\mu^0)$ precedes $\lambda^1(\mu)$, and for any (μ) , $\lambda^1(\mu)$ precedes $\lambda^i(\mu)$ if $i > 1$. Then from (43) and (37) we get (since the γ 's and δ 's are non-negative integers),

LEMMA VI. Let $(\lambda^0) = \lambda^1(\mu^0) = (\mu_1^0 + p, \mu_2^0, \dots)$. Then (i) if (μ^0) precedes (μ) $\mathfrak{T}_{(\lambda^0)}$ contains no constituents from the block $\mathfrak{B}(\mu)$, and (ii) $\mathfrak{F}_1(\mu^0)$ is the only constituent of $\mathfrak{T}_{(\lambda^0)}$ belonging to the block $\mathfrak{B}(\mu^0)$ and it occurs with multiplicity one (i.e. $\gamma_0(\mu^0 \check{\lambda}^0) = 1$).

In §6 we saw that the commutator algebra of a representation $\mathfrak{B} = \sum_p \mathfrak{T}_{(\lambda^p)}$ can be computed block at a time. Let $N(\mathfrak{B}, (\mu))$ denote the contribution to $N(\mathfrak{B})$ of a block of lowest kind and $N(\mathfrak{B}, (\lambda))$ the same for a block of highest kind. Then (see formula (32)) we have

$$(44) \quad \begin{aligned} N(\mathfrak{B}, (\mu)) &= \sum_{i=1}^p (\sum_p \gamma_{i-1}(\mu \check{\lambda}^p) + \gamma_i(\mu \check{\lambda}^p))^2; \\ N(\mathfrak{B}, (\lambda)) &= (\sum_p \delta(\lambda \check{\lambda}^p))^2 \end{aligned}$$

and for the total order

$$(45) \quad N(\mathfrak{B}) = \sum N(\mathfrak{B}, (\mu)) + \sum N(\mathfrak{B}, (\lambda)),$$

⁸ [2], pp. 71, 94; [3], p. 110; [6], p. 205.

where the first sum is over all partitions (μ) of l and the second sum is over all partitions (λ) of m whose diagrams have no hook of length p .

For the corresponding non-modular representation \mathfrak{B}^0 we have analogously

$$N(\mathfrak{B}^0, (\lambda)) = (\sum_p \delta(\lambda \check{\chi} \lambda^p))^2,$$

and $N(\mathfrak{B}^0) = \sum N(\mathfrak{B}^0, (\lambda))$, the sum extending over all partitions (λ) of m . For comparison with $N(\mathfrak{B})$ it is convenient to group together those $N(\mathfrak{B}^0, (\lambda))$ for which the representations $\mathfrak{F}(\lambda)$ belong to the same blocks. Thus we obtain

$$(46) \quad N(\mathfrak{B}^0) = \sum N(\mathfrak{B}^0, (\mu)) + \sum N(\mathfrak{B}^0, (\lambda))$$

where the summation ranges are the same as in (45) and

$$(47) \quad N(\mathfrak{B}^0, (\mu)) = \sum_{i=1}^p N(\mathfrak{B}^0, \lambda^i(\mu)) = \sum_{i=1}^p (\sum_p \delta_i(\mu \check{\chi} \lambda^p))^2.$$

By Remark I the difference $N(\mathfrak{B}^0) - N(\mathfrak{B})$ is zero. Observe that the $N(\mathfrak{B}^0, (\lambda))$ and $N(\mathfrak{B}, (\lambda))$ from (45) and (46) cancel. Furthermore, by our induction hypothesis, (40) holds for any (μ) which precedes (μ^0) ; this in turn implies that $N(\mathfrak{B}^0, (\mu)) = N(\mathfrak{B}, (\mu))$ for any (μ) which precedes (μ^0) . Hence we have

$$(48) \quad 0 = \sum N(\mathfrak{B}^0, (\mu)) - N(\mathfrak{B}, (\mu))$$

where the summation extends over all (μ) (including (μ^0)) which do not precede (μ^0) .

Now let (λ) be any partition of m . By (43), (47), $N(\mathfrak{T}_{(\lambda)}^0 + \mathfrak{T}_{(\lambda^0)}^0, (\mu)) = N(\mathfrak{T}_{(\lambda)}^0, (\mu))$ if (μ) follows (μ^0) . By Lemma VI and (44), $N(\mathfrak{T}_{(\lambda)}^0 + \mathfrak{T}_{(\lambda^0)}^0, (\mu)) = N(\mathfrak{T}_{(\lambda)}^0, (\mu))$ if (μ) follows (μ^0) . Hence if we subtract (48) for $\mathfrak{B} = \mathfrak{T}_{(\lambda)}$ from (48) for $\mathfrak{B} = \mathfrak{T}_{(\lambda)} + \mathfrak{T}_{(\lambda^0)}$ the only terms which do not cancel are those involving (μ^0) . This leads us to

$$(49) \quad \begin{aligned} N(\mathfrak{T}_{(\lambda)}^0 + \mathfrak{T}_{(\lambda^0)}^0, (\mu^0)) - N(\mathfrak{T}_{(\lambda)}^0, (\mu^0)) \\ = N(\mathfrak{T}_{(\lambda)} + \mathfrak{T}_{(\lambda^0)}, (\mu^0)) - N(\mathfrak{T}_{(\lambda)}, (\mu^0)). \end{aligned}$$

Now substituting in this from (44) and (47) and referring to (43) and Lemma VI for the values of $\delta_i(\mu^0 \check{\chi} \lambda^0)$, $\gamma_i(\mu^0 \check{\chi} \lambda^0)$ we have

$$\begin{aligned} & [(\delta_1(\mu^0 \check{\chi} \lambda) + 1)^2 + \delta_2(\mu^0 \check{\chi} \lambda)^2 + \cdots + \delta_p(\mu^0 \check{\chi} \lambda)^2] - [\delta_1(\mu^0 \check{\chi} \lambda)^2 + \cdots + \delta_p(\mu^0 \check{\chi} \lambda)^2] \\ & = [(\gamma_0(\mu^0 \check{\chi} \lambda) + \gamma_1(\mu^0 \check{\chi} \lambda) + 1)^2 + (\gamma_1(\mu^0 \check{\chi} \lambda) + \gamma_2(\mu^0 \check{\chi} \lambda))^2 + \cdots + \gamma_{p-1}(\mu^0 \check{\chi} \lambda)^2] \\ & \quad - [(\gamma_0(\mu^0 \check{\chi} \lambda) + \gamma_1(\mu^0 \check{\chi} \lambda))^2 + \cdots + \gamma_{p-1}(\mu^0 \check{\chi} \lambda)^2] \end{aligned}$$

or

$$2\delta_1(\mu^0 \check{\chi} \lambda) + 1 = 2(\gamma_0(\mu^0 \check{\chi} \lambda) + \gamma_1(\mu^0 \check{\chi} \lambda)) + 1$$

which is the same as (40). Observe that the argument above applies to the first partition, $\mu_1 = l$ of l ; hence our induction is complete and Theorem VI is fully established.

In words (49) says that the change in order of the commutator algebra, due to addition of one particular permutation representation to another particular permutation representation, is the same, block by block, characteristic 0 as characteristic p . Remark I states merely that the total change is the same. If we could strengthen Remark I to a block by block form, the proof of the above theorem could be much shortened, as then one could omit everything between Lemma VI and formula (49); and of course any such improvement of Remark I would be of interest in the general modular theory entirely aside from its application here.

8. The Kronecker m^{th} power of the full linear group

In order to describe the structure of the enveloping algebra \mathfrak{A}_m of Π_m we have only to put together the results of the preceding sections and introduce a suitable notation for the constituents.

Let (λ) be a partition of m whose diagram has no hook of length p . Then we denote by $\mathfrak{G}(\lambda)$ the total \mathfrak{k} -matrix algebra of degree $\delta(\lambda)$. Let (μ) be a partition of $l = m - p$. Then we denote by $\mathfrak{G}_i(\mu)$ the total \mathfrak{k} -matrix algebra of degree $^9 \gamma_i(\mu)$, and by $\mathfrak{G}_j^i(\mu)$, for $j = i - 1, i, i + 1$, the set of all $\gamma_i(\mu)$ by $\gamma_j(\mu)$ \mathfrak{k} -matrices; all this for $i = 0, \dots, p - 1$, with the usual conventions for $i = 0, i = p - 1$ (i.e., $j < 0, j > p - 1$ are excluded): Finally, we set

$$(50) \quad \mathfrak{U}^i(\mu) = \left\| \begin{array}{cccc} \mathfrak{G}_i(\mu) & & & \\ \mathfrak{G}_i^{i-1}(\mu) & \mathfrak{G}_{i-1}(\mu) & & \\ \mathfrak{G}_i^{i+1}(\mu) & 0 & \mathfrak{G}_{i+1}(\mu) & \\ \mathfrak{G}_i^i(\mu) & \mathfrak{G}_{i-1}^i(\mu) & \mathfrak{G}_{i+1}^i(\mu) & \mathfrak{G}_i(\mu) \end{array} \right\| \quad i = 1, \dots, p - 1.$$

We can now state the main theorem on the structure of \mathfrak{A}_m .

THEOREM VII. *For $m < 2p$ and \mathfrak{k} any field (of characteristic p) containing at least m elements, the enveloping algebra \mathfrak{A}_m , of the Kronecker m^{th} power representation Π_m , of the full linear group \mathfrak{G} , of n -rowed non-singular \mathfrak{k} -matrices, has the following indecomposable (direct) constituents: (i) If (λ) is any partition of m whose diagram has no hook of length p , then $\mathfrak{G}(\lambda)$ appears $f(\lambda)$ times as an indecomposable constituents of \mathfrak{A}_m ; where $f(\lambda)$ is the degree of the ordinary irreducible representation of \mathfrak{S}_m defined by (λ) . (ii) If (μ) is any partition of $l = m - p$, then $\mathfrak{U}^i(\mu)$ appears $f_i(\mu)$ times as an indecomposable constituent of \mathfrak{A}_m , $i = 1, \dots, p - 1$, where $f_i(\mu)$ is the degree of the irreducible modular representation $\mathfrak{F}_i(\mu)$ of \mathfrak{S}_m .*

The theorem follows from Theorem III, the formulas of §6, and Theorem VI. The degrees $f(\lambda)$ are well known [6, p. 213], and for the $^{10} f_i(\mu)$ we have

$$(51) \quad f_i(\mu) = f(\lambda^i(\mu)) - f(\lambda^{i-1}(\mu)) + \dots + (-1)^{i+1} f(\lambda^1(\mu))$$

⁹ See formula (38) for the value of $\gamma_i(\mu)$.

¹⁰ Cf. [5], formula (20).

where¹¹ $\lambda^j(\mu)$ is the partition of m whose diagram T has a hook H , of length p and height (vertical length) j , such that $T - H$ is the diagram of (μ) .

Any indecomposable constituent of \mathfrak{A}_m affords an indecomposable representation of \mathfrak{G} . We shall consider the symbols $\mathfrak{G}(\lambda)$, $\mathfrak{U}^i(\mu)$, $\mathfrak{G}_i(\mu)$ in two ways: first, as they are defined above; and second, as denoting representations of \mathfrak{G} ; for instance $\mathfrak{G}(\lambda)$ is the representation $A \rightarrow G(\lambda\chi A)$, where $G(\lambda\chi A)$ is the matrix in $\mathfrak{G}(\lambda)$ assigned to $\Pi_m(A)$ considered as an element of \mathfrak{A}_m . We define the matrices $U^i(\mu\chi A)$, $G_i^i(\mu\chi A)$ analogously. Then the $\mathfrak{G}(\lambda)$, $\mathfrak{G}_i(\mu)$ are all the irreducible representations of \mathfrak{G} which are induced in the space of tensors of rank m .

INSTITUTE FOR ADVANCED STUDY

BIBLIOGRAPHY

- [1] M. DEURING, *Algebren*, Berlin, 1935.
- [2] D. E. LITTLEWOOD, *The Theory of Group Characters*, Oxford, 1940.
- [3] F. MURNAGHAN, *The Theory of Group Representations*, Baltimore, 1938.
- [4] C. J. NESBITT, *On the Regular Representations of Algebras*, these Annals, vol. 39 (1938), pp. 634-658.
- [5] R. M. THRALL AND C. J. NESBITT, *On the Modular Representations of the Symmetric Group*, these Annals, vol. 43 (1942), pp. 656-670.
- [6] H. WEYL, *The Classical Groups*, Princeton, 1939.

¹¹ See [5], §6, for a discussion of hooks and for references to Nakayama's treatment.

NON-ASSOCIATIVE ALGEBRAS¹

I. Fundamental Concepts and Isotopy

BY A. A. ALBERT

(Received January 5, 1942)

1. Introduction

The study of non-associative algebras has already yielded much of interest and importance. Indeed those special theories² in which the associative law is replaced by a substitute have each been of an extent and of an interest almost comparable to that of the theory of associative algebras.

The results on non-associative algebras in which one does not assume a type of partial associativity³ have almost⁴ all been of a rather primitive kind and have been scattered through the literature. They have, in particular, not emphasized adequately the important fact that many of the properties of arbitrary linear algebras are equivalent to certain properties of related sets of linear transformations in which multiplication then does satisfy the associative law. The fact that there is a rather surprisingly large number of non-associative algebras of orders two and three has been noted³ but it has not been recognized before that this is at least partly due to the undesirable narrowness of the concept of equivalence for algebras other than associative algebras with a unity quantity.

It is the purpose of that part of the study of non-associative algebras which we shall begin here to emphasize the facts noted above by providing an appropriate formulation of the fundamentals of the theory of arbitrary linear algebras. Thus we shall devote the first portion of our present discussion to the process of relating the elementary properties of an algebra \mathfrak{A} to the corresponding properties of three attached linear spaces of linear transformations on \mathfrak{A} . We shall then introduce the concept of *isotopy* of algebras, an extension of the concept of equivalence which coincides with the latter concept in the case of associative algebras with a unity quantity. Our discussion will conclude with an extensive

¹ Presented to the Society September 5, 1941. Most of the results of this paper were announced also in lectures at Princeton and Harvard in March 1941.

² We refer here first to the theory of alternative algebras for which see M. Zorn, *Theorie der Alternative Ringe*, Hamburg Abh., vol. 8 (1930), pp. 123-47, *Alternativkörper und Quadratische Systeme*, loc. cit., vol. 9 (1933), pp. 395-402. A second such theory is that of Lie Algebras for which see N. Jacobson, *Simple Lie algebras over a field of characteristic zero*, Duke J., vol. 4 (1938), pp. 534-51. Finally Jordan algebras are described in P. Jordan, J. v. Neumann, and E. Wigner, *On an algebraic generalization of the quantum mechanical formalism*, these Annals, vol. 35 (1934), pp. 29-64. The articles quoted contain bibliographies of their subjects and it should be remarked here that the theory of Jordan algebras has been generalized by G. Kalisch in his Chicago doctoral dissertation.

³ Cf. L. E. Dickson, *Linear algebras with associativity not assumed*, Duke J., vol. 1 (1935), pp. 113-25.

⁴ For a paper not of this type see footnote 6.

consideration of the question as to what properties of an algebra are preserved when we pass to an isotope.

It is the author's hope that the study begun here may ultimately lead to a solution of the problem of determining all simple algebras with a unity quantity, at least in a sense like that in which we say that the corresponding problem for associative algebras has been solved. A fundamental part of the associative algebra theory consisted of the definition of the known types of algebras, that is, the cyclic algebras and the crossed products, and such definitions will also be required for the non-associative case. We shall provide at least a very extensive part of this requirement in Part II of the present study. There we shall define⁵ a class of non-commutative simple algebras with a unity quantity containing all such algebras which have been considered thus far in the literature as well as a very rich variety of new types.

2. The multiplication spaces of an algebra

A linear algebra of order n over a field \mathfrak{F} is, in particular, a linear space of order n over \mathfrak{F} . But all linear spaces of the same order are equivalent. Thus we may regard all algebras of the same order as having the same quantities but with different laws for forming products. In particular we may take our quantities to be vectors, that is, one by n matrices

$$(1) \quad a = (\alpha_1, \dots, \alpha_n) \quad (\alpha_i \text{ in } \mathfrak{F}).$$

An algebra \mathfrak{A} now consists of the linear space \mathfrak{L} of all the vectors (1) together with a set of n^3 quantities γ_{ijk} in \mathfrak{F} such that the product

$$(2) \quad u = a \cdot x = (\alpha_1, \dots, \alpha_n) \cdot (\xi_1, \dots, \xi_n) = (\mu_1, \dots, \mu_n)$$

of any two quantities of \mathfrak{L} is defined in \mathfrak{A} by

$$(3) \quad \mu_k = \sum_{i,j=1}^n \alpha_i \gamma_{ijk} \xi_j \quad (k = 1, \dots, n).$$

Define $\Gamma^{(j)}$ to be the n -rowed square matrix with γ_{ijk} in the i^{th} row and k^{th} column, and write

$$(4) \quad \Gamma_x = \Gamma^{(1)}\xi_1 + \dots + \Gamma^{(n)}\xi_n,$$

so that $\Gamma^{(i)} = \Gamma_{e_i}$ where e_i is given by (1) with $\alpha_i = 1$ and all the other $\alpha_j = 0$. The customary row by column definition of a matrix product does not include the definition of ax and so $a \cdot x \neq ax$. However it is clear that

$$(5) \quad a \cdot x = a\Gamma_x,$$

where $a\Gamma_x$ is computed as usual. Matrix multiplication is associative and so

$$(6) \quad (a \cdot x) \cdot y = (a\Gamma_x)\Gamma_y = a(\Gamma_x\Gamma_y).$$

⁵ This definition has already been presented by the author in a lecture at the University of Cincinnati, November 15, 1941. See also the author's paper on *Quadratic forms permitting composition*, these Annals, this volume.

But

$$(7) \quad a \cdot (x \cdot y) = a\Gamma_u, \quad u = x \cdot y = x\Gamma_y,$$

and \mathfrak{A} is associative if and only if $\Gamma_x\Gamma_y = \Gamma_u$ for every x and y . We shall obtain another criterion later.

The linear transformation R_x on \mathfrak{L} whose matrix is Γ_x is the correspondence

$$(8) \quad a \rightarrow aR_x = a \cdot x,$$

and is called a *right multiplication* of \mathfrak{A} . Its matrix Γ_x depends upon our choice of the linear space equivalence between \mathfrak{A} and \mathfrak{L} . However the transformation R_x does not depend upon this choice. We shall therefore study the properties of R_x rather than of Γ_x . However our present notation aR_x for the result of applying R_x to a has the advantage that $aR_x = a\Gamma_x$, so that in computations we may replace R_x by its matrix. We shall not use the author's earlier notation a^{R_x} .

The set

$$(9) \quad R(\mathfrak{A})$$

of all right multiplications of \mathfrak{A} is a linear subspace of order at most n of the total matrix algebra $(\mathfrak{F})_n$ (of order n^2 over \mathfrak{F}) of all linear transformations on \mathfrak{L} . The correspondence

$$(10) \quad x \rightarrow R_x$$

is a linear mapping of \mathfrak{L} on $R(\mathfrak{A})$, and $R(\mathfrak{A})$ is spanned by R_{e_1}, \dots, R_{e_n} .

We may now state that any algebra \mathfrak{A} of order n over \mathfrak{F} consists of a linear space \mathfrak{L} of this same order, a linear space $R(\mathfrak{A})$ of order $m \leq n$ over \mathfrak{F} consisting of linear transformations on \mathfrak{L} , and a linear mapping (10) of \mathfrak{L} on $R(\mathfrak{A})$. Conversely let \mathfrak{L} and a subspace \mathfrak{N} of $(\mathfrak{F})_n$ be given such that the order of \mathfrak{N} is at most n . Then we may select any n transformations $R^{(i)}$ which span \mathfrak{N} and define $R_x = R^{(1)}\xi_1 + \dots + R^{(n)}\xi_n$. This then determines a linear mapping of \mathfrak{L} on \mathfrak{N} and hence an algebra \mathfrak{A} with $\mathfrak{N} = R(\mathfrak{A})$. It is particularly important to note that \mathfrak{N} is completely arbitrary save for the upper bound n on its order.

The linear transformations L_x given by

$$(11) \quad a \rightarrow x \cdot a = aL_x$$

are called *left multiplications* of \mathfrak{A} and form the *left multiplication space* $L(\mathfrak{A})$ of \mathfrak{A} . This space and the linear mapping

$$(12) \quad x \rightarrow L_x$$

of \mathfrak{L} on $L(\mathfrak{A})$ determine and are completely determined by $R(\mathfrak{A})$ and (10). For the matrix of L_x is $\Delta_x = \Delta^{(1)}\xi_1 + \dots + \Delta^{(n)}\xi_n$, where $\Delta^{(i)}$ is the matrix with $\gamma_{i,jk}$ in the j^{th} row and k^{th} column. Correspondingly $L^{(1)}\xi_1 + \dots + L^{(n)}\xi_n = L_x$ so that $L(\mathfrak{A})$ is spanned over \mathfrak{F} by $L^{(1)}, \dots, L^{(n)}$.

The right and left multiplication spaces of \mathfrak{A} generate another linear sub-

space of $(\mathfrak{F})_n$ which we shall call the *transformation algebra* of \mathfrak{A} . It is the algebra

$$(13) \quad T(\mathfrak{A}) = \mathfrak{F}[I, R^{(1)}, \dots, R^{(n)}, L^{(1)}, \dots, L^{(n)}]$$

of all polynomials with coefficients in \mathfrak{F} in the $R^{(i)}$ which span $R(\mathfrak{A})$, the $L^{(i)}$ which span $L(\mathfrak{A})$ and the identity transformation I of $(\mathfrak{F})_n$. All three spaces $R(\mathfrak{A})$, $L(\mathfrak{A})$, $T(\mathfrak{A})$ will be used to describe properties of \mathfrak{A} and we shall call them the *multiplication spaces* of \mathfrak{A} .

The scalar extension $\mathfrak{A}_{\mathfrak{K}}$ of \mathfrak{A} by any scalar extension field \mathfrak{K} of \mathfrak{F} is the set of vectors (1) with the α_i in \mathfrak{K} and with $a \cdot x$ defined in $\mathfrak{A}_{\mathfrak{K}}$ by the same γ_{ij} as define \mathfrak{A} . Then clearly we have the same $R^{(i)}$ and $L^{(i)}$, and

$$(14) \quad R(\mathfrak{A}_{\mathfrak{K}}) = [R(\mathfrak{A})]_{\mathfrak{K}}, \quad L(\mathfrak{A}_{\mathfrak{K}}) = [L(\mathfrak{A})]_{\mathfrak{K}}, \quad T(\mathfrak{A}_{\mathfrak{K}}) = [T(\mathfrak{A})]_{\mathfrak{K}}.$$

When we begin to discuss more than one algebra \mathfrak{A} defined for the same \mathfrak{F} it will be necessary to distinguish \mathfrak{A} from the fixed linear space \mathfrak{L} of the quantities of \mathfrak{A} . However no confusion will arise if we speak of a linear subspace of \mathfrak{L} as a *subspace* of \mathfrak{A} and this will be desirable, of course, in discussing subalgebras of \mathfrak{A} .

3. Products of spaces

In our study of the multiplication spaces of an algebra we shall need to use the notations for products of spaces of both the algebra \mathfrak{A} and the algebra $(\mathfrak{F})_n$. We define the *product*

$$\mathfrak{B}\mathfrak{C}$$

of any two linear subspaces of an algebra \mathfrak{A} to be the linear subspace over \mathfrak{F} of \mathfrak{A} spanned by $b_i \cdot c_j$ ($i = 1, \dots, s; j = 1, \dots, t$), where b_1, \dots, b_s span \mathfrak{B} and c_1, \dots, c_t span \mathfrak{C} . Then the square $\mathfrak{B}^2 = \mathfrak{B}\mathfrak{B}$ is defined and we define the right power $\mathfrak{B}^{k+1} = \mathfrak{B}^k\mathfrak{B}$.

If a is in \mathfrak{A} the subspace (of order zero or one over \mathfrak{F}) of \mathfrak{A} spanned by a will be designated by $a\mathfrak{F}$. Then we write $a\mathfrak{B}$ for $(a\mathfrak{F})\mathfrak{B}$ and similarly $\mathfrak{B}a$ for $\mathfrak{B}(a\mathfrak{F})$. If c is also in \mathfrak{A} we write $(a\mathfrak{B})c$ for $(a\mathfrak{B})(c\mathfrak{F})$ and $a(\mathfrak{B}c)$ for $(a\mathfrak{F})(\mathfrak{B}c)$. If \mathfrak{A} is associative and $\mathfrak{B}, \mathfrak{C}, \mathfrak{D}$ are linear subspaces then $(\mathfrak{B}\mathfrak{C})\mathfrak{D} = \mathfrak{B}(\mathfrak{C}\mathfrak{D})$, $b\mathfrak{C}d$ is defined to be $(b\mathfrak{C})d = b(\mathfrak{C}d)$ for every b and d in \mathfrak{A} .

The definitions above apply of course both to subspaces of any algebra \mathfrak{A} of order n over \mathfrak{F} and to subspaces of $(\mathfrak{F})_n$. However let \mathfrak{B} be a linear subspace of \mathfrak{A} and \mathfrak{C} be a linear subspace of $(\mathfrak{F})_n$. We define

$$(15) \quad \mathfrak{B}\mathfrak{C}$$

to be the linear subspace of \mathfrak{A} spanned over \mathfrak{F} by the products bS for b in \mathfrak{B} and S in \mathfrak{C} . Then we have defined the *product operation* (15) as an operation on $\mathfrak{A}, (\mathfrak{F})_n$ to \mathfrak{A} . We also define $b\mathfrak{C} = (b\mathfrak{F})\mathfrak{C}$, $\mathfrak{B}S = \mathfrak{B}(S\mathfrak{F})$ for all linear subspaces \mathfrak{B} of \mathfrak{A} and \mathfrak{C} of $(\mathfrak{F})_n$, and all quantities b in \mathfrak{A} and S in $(\mathfrak{F})_n$. Clearly $(\mathfrak{B}\mathfrak{C})\mathfrak{I} = \mathfrak{B}(\mathfrak{C}\mathfrak{I})$ for all linear subspaces \mathfrak{B} of \mathfrak{A} and \mathfrak{C} and \mathfrak{I} of $(\mathfrak{F})_n$. We now prove N. Jacobson's

LEMMA 1. Let \mathfrak{N} be a nilpotent subalgebra of $(\mathfrak{F})_n$. Then $\mathfrak{N}\mathfrak{N} \neq \mathfrak{A}$ or zero.

For $\mathfrak{N} \neq 0$ and contains an $S \neq 0$ of $(\mathfrak{F})_n$, $aS \neq 0$ for some a of \mathfrak{A} and is in $\mathfrak{N}\mathfrak{N} \neq 0$. If $\mathfrak{N}\mathfrak{N} = \mathfrak{A}$ then $\mathfrak{N}\mathfrak{N}^2 = (\mathfrak{N}\mathfrak{N})\mathfrak{N} = \mathfrak{N}\mathfrak{N} = \mathfrak{A}$ and $\mathfrak{N}\mathfrak{N}^k = \mathfrak{A}$ implies that $\mathfrak{N}\mathfrak{N}^{k+1} = (\mathfrak{N}\mathfrak{N}^k)\mathfrak{N} = \mathfrak{N}\mathfrak{N} = \mathfrak{A}$. Hence $\mathfrak{N}\mathfrak{N}^t = \mathfrak{A}$ for every t . But $\mathfrak{N}^t = 0$ for some t , $\mathfrak{A} = 0$ which is impossible.

If E is in $(\mathfrak{F})_n$ and \mathfrak{S} is a linear subspace of $(\mathfrak{F})_n$ the condition $E\mathfrak{S} = E\mathfrak{S}E$ means that $ES = EUE$ for every S of \mathfrak{S} where U in \mathfrak{S} is determined (but not necessarily uniquely) by S . However when E is an idempotent, that is $E^2 = E$, the property $E\mathfrak{S} = E\mathfrak{S}E$ is equivalent to $ES = ESE$ for every S of \mathfrak{S} . For $ES = EUE = EUE^2 = (EUE)E = ESE$.

4. Subalgebras

If \mathfrak{B} is a linear subspace of an algebra \mathfrak{A} the set of all R_y for y in \mathfrak{B} is a linear subspace of $R(\mathfrak{A})$, the set of all L_y is a linear subspace of $L(\mathfrak{A})$, and these subspaces, together with I , generate a subalgebra of $T(\mathfrak{A})$. We designate these three linear subspaces of $T(\mathfrak{A})$ by

$$(16) \quad R(\mathfrak{B}, \mathfrak{A}), \quad L(\mathfrak{B}, \mathfrak{A}), \quad T(\mathfrak{B}, \mathfrak{A})$$

respectively, where $T(\mathfrak{B}, \mathfrak{A})$ is the set of all polynomials with coefficients in \mathfrak{F} in the R_y , the L_y and I .

If \mathfrak{B} has order $m < n$ over \mathfrak{F} we may express \mathfrak{A} as the supplementary sum $\mathfrak{B} + \mathfrak{C}$ where \mathfrak{C} has order $n - m$. This means that every a of \mathfrak{A} is uniquely expressible in the form $a = b + c$ for b in \mathfrak{B} and c in \mathfrak{C} . However \mathfrak{C} is by no means unique. We now define a mapping

$$E: \quad a = b + c \rightarrow b = aE$$

of \mathfrak{A} on \mathfrak{B} . It is an idempotent linear transformation of rank m , that is, $E^2 = E$ and m is the rank of the matrix of E . We then have $\mathfrak{B} = \mathfrak{A}E$ where E is characterized by the property that $a = aE$ if and only if a is in \mathfrak{B} , $aE = 0$ if and only if a is in \mathfrak{C} . A corresponding idempotent for \mathfrak{C} is $I - E$. We now have

LEMMA 2. Let \mathfrak{B} be a linear subspace of order m of \mathfrak{A} so that $\mathfrak{B} = \mathfrak{A}E$ for an idempotent E of rank m in $(\mathfrak{F})_n$. Then \mathfrak{B} is a subalgebra of \mathfrak{A} if and only if $E[R(\mathfrak{B}, \mathfrak{A})] = E[R(\mathfrak{B}, \mathfrak{A})]E$.

For \mathfrak{B} is a subalgebra of \mathfrak{A} if and only if $aE \cdot y = (aE \cdot y)E$ for every a of \mathfrak{A} and y of \mathfrak{B} . Then $aER_y = aER_yE$, $ER_y = ER_yE$ and we have our lemma since $E^2 = E$.

Note that also $y \cdot aE = aEL_y = aEL_yE$, $E[L(\mathfrak{B}, \mathfrak{A})] = E[L(\mathfrak{B}, \mathfrak{A})]E$. Since $EIE = E = EI$ we see that $EU = EUE$ for every U of IF , $R(\mathfrak{B}, \mathfrak{A})$, $L(\mathfrak{B}, \mathfrak{A})$. But if $ES = ESE$ and $EU = EUE$ we have $E(S + U) = E(S + U)E$, $ESU = ESEU = ESEUE = ESUE$. Hence $E[T(\mathfrak{B}, \mathfrak{A})] = E[T(\mathfrak{B}, \mathfrak{A})]E$. The converse is trivial and we have

LEMMA 2'. Let $\mathfrak{B} = \mathfrak{A}E$ as in Lemma 2. Then \mathfrak{B} is a subalgebra of \mathfrak{A} if and only if $E[T(\mathfrak{B}, \mathfrak{A})] = E[T(\mathfrak{B}, \mathfrak{A})]E$.

5. Ideals

A subspace \mathfrak{B} of an algebra \mathfrak{A} is a right ideal of \mathfrak{A} if and only if $y \cdot x$ is in \mathfrak{B} for every y of \mathfrak{B} and x of \mathfrak{A} . Then $\mathfrak{B} = \mathfrak{A}E$ for an idempotent E of rank equal to the order of \mathfrak{B} over \mathfrak{F} and \mathfrak{B} is a right ideal of \mathfrak{A} if and only if either of the following conditions

$$(17) \quad L_y = L_y E, \quad ER_x = ER_x E \quad (x \text{ in } \mathfrak{A}, y = yE \text{ in } \mathfrak{B})$$

holds. For $y \cdot x = xL_y = xL_y E$, $y = aE$, $aE \cdot x = aER_x = aER_x E$. We may state this result as

LEMMA 3. *Let $\mathfrak{B} = \mathfrak{A}E$ for an idempotent E of \mathfrak{A} . Then \mathfrak{B} is a right ideal of \mathfrak{A} if and only if $ER(\mathfrak{A}) = ER(\mathfrak{A})E$. This is equivalent to the condition that $L(\mathfrak{B}, \mathfrak{A})$ be contained in $[L(\mathfrak{A})]E$.*

In the theory of group representations the property $ER(\mathfrak{A}) = ER(\mathfrak{A})E$ for $E \neq 0$, I is called the property that $R(\mathfrak{A})$ is a *reducible set* of linear transformations. We shall not use this terminology again here.

Left ideals are defined similarly and $\mathfrak{B} = \mathfrak{A}E$ is a left ideal if and only if $EL(\mathfrak{A}) = EL(\mathfrak{A})E$, and thus if and only if $R(\mathfrak{B}, \mathfrak{A})$ is in $[R(\mathfrak{A})]E$. We call \mathfrak{B} an ideal of \mathfrak{A} if it is both a left and a right ideal. This occurs if and only if $\mathfrak{B} = \mathfrak{A}E$ where $EU = EUE$ for every U in either $R(\mathfrak{A})$ or $L(\mathfrak{A})$. As in the proof of Lemma 2 we have $EU = EUE$ for every U of $T(\mathfrak{A})$ and have

LEMMA 4. *A linear subspace $\mathfrak{B} = \mathfrak{A}E$ of \mathfrak{A} is an ideal of \mathfrak{A} if and only if $ET(\mathfrak{A}) = ET(\mathfrak{A})E$.*

We shall call a quantity a of \mathfrak{A} *right singular* or *right non-singular* according as R_a is or is not singular. We then have

LEMMA 5. *Let $\mathfrak{B} = \mathfrak{A}E$ be an ideal of \mathfrak{A} and a be a right non-singular quantity of \mathfrak{A} . Then $E(R_a)^{-1} = E(R_a)^{-1}E$.*

For R_a is in the associative algebra $T(\mathfrak{A})$ and so is $(R_a)^{-1}$, $ET(\mathfrak{A}) = ET(\mathfrak{A})E$.

We next prove

LEMMA 6. *Let P be in $(\mathfrak{F})_n$ and $\mathfrak{F}[P]$ be a field of degree n over \mathfrak{F} . Then $EP = EPE$ for an idempotent E of $(\mathfrak{F})_n$ if and only if $E = I$ or $E = 0$.*

For let $EP = EPE$ and $E \neq 0$. Then EP is in the total matrix algebra $E(\mathfrak{F})_n E$ with unity quantity E , a total matrix algebra whose degree m is the rank of E . If $EP^k = EP^k E$ then $EP^{k+1} = EP^k EP = EP^k EPE = EP^{k+1} E$, $EP^t = EP^t E$ for every t . But then $(EP)^t = EP^t E$ since from $(EP)^k = EP^k E$ we have $(EP)^{k+1} = EP^k EEP = EP^{k+1} E$. It follows that $\phi(EP) = E\phi(P)E$ for any polynomial $\phi(P)$. But if $\phi(\lambda)$ is the minimum function of P it is an irreducible polynomial of degree n and $\phi(EP) = 0$. This is impossible when E is singular since the minimum function of EP in a total matrix algebra of degree $m < n$ has degree at most m . Thus $E = I$.

6. Divisors of zero

If b is right non-singular the equation $x \cdot b = a$ has the unique solution $x = a(R_b)^{-1}$. However there exists a $c \neq 0$ such that $c \cdot b = 0$ when b is right singular. We shall call b a *right divisor of zero* if it is a non-zero right singular

quantity. *Left singularity* and *left non-singularity* as well as *left divisors of zero* are defined similarly, and it is clear that an algebra contains right divisors of zero b if and only if it contains left divisors of zero c .

A quantity b of an algebra \mathfrak{A} is called an *absolute right divisor of zero* if $b \neq 0$ and $a \cdot b = 0$ for every a of \mathfrak{A} . But then $R_b = 0$. This can occur only if the linear mapping $x \rightarrow R_x$ of \mathfrak{A} on $R(\mathfrak{A})$ is singular, that is, if and only if $R(\mathfrak{A})$ has smaller order than \mathfrak{A} . Similarly $L(\mathfrak{A})$ has order less than the order of \mathfrak{A} if and only if some quantity b in \mathfrak{A} is an absolute left divisor of zero, that is, $L_b = 0$ and $b \neq 0$. Each absolute right (left) divisor of zero spans a subalgebra $\mathfrak{B} = b\mathfrak{F}$ of \mathfrak{A} which is a zero algebra of order one and is a left (right) ideal of \mathfrak{A} . In fact we have

LEMMA 7. A linear subspace $\mathfrak{B} = \mathfrak{A}E = b\mathfrak{F}$ for an absolute right divisor of zero b of \mathfrak{A} if and only if $EL(\mathfrak{A}) = 0$, E has rank one.

For $\mathfrak{B} = b\mathfrak{F} = \mathfrak{A}E$ where E has rank one. Then aE is zero or an absolute right divisor of zero if and only if $x \cdot aE = aEL_x = 0$, $EL_x = 0$, $EL(\mathfrak{A}) = 0$.

If $T(\mathfrak{A}) = I\mathfrak{F}$ then $R_a = \alpha I$, $L_a = \beta I$ for every a of \mathfrak{A} where α and β are in \mathfrak{F} . If $R_a \neq 0$ for some a then $x \cdot a = aL_x = xR_a = x\alpha \neq 0$ and $L_x \neq 0$. It follows that the mapping $x \rightarrow L_x$ is one-to-one, $n = 1$. Similarly if $L_a \neq 0$ we have $n = 1$. Thus $n > 1$ implies that $T(\mathfrak{A}) = I\mathfrak{F}$ only if $R(\mathfrak{A}) = L(\mathfrak{A}) = 0$, \mathfrak{A} is a zero algebra. The converse is trivial and we have

LEMMA 8. The algebra $T(\mathfrak{A}) = I\mathfrak{F}$ if and only if \mathfrak{A} is a zero algebra or $n = 1$ and $\mathfrak{A} = \mathfrak{F}$.

We call b an *absolute divisor of zero* if it is both an absolute right and an absolute left divisor of zero. For algebras containing such quantities we have

LEMMA 9. An algebra \mathfrak{A} contains absolute divisors of zero if and only if there exists a non-zero idempotent E of $(\mathfrak{F})_n$ such that $ER(\mathfrak{A}) = EL(\mathfrak{A}) = 0$. Then

$$(18) \quad T(\mathfrak{A}) = \mathfrak{S} + I\mathfrak{F}$$

where \mathfrak{S} is the set of all transformations S of $T(\mathfrak{A})$ such that $ES = 0$ and is an ideal of $T(\mathfrak{A})$.

For if b is an absolute divisor of zero the proof of Lemma 7 implies that there exists a non-zero idempotent E such that $ER(\mathfrak{A}) = EL(\mathfrak{A}) = 0$. Conversely if $ER(\mathfrak{A}) = EL(\mathfrak{A}) = 0$ the quantities of $\mathfrak{A}E$ have the property $aE \cdot x = aER_x = 0 = aEL_x = x \cdot aE$. Then $\mathfrak{A}E \neq 0$ consists of zero and absolute divisors of zero. The quantities of $T(\mathfrak{A})$ are sums of scalars αI for α in \mathfrak{F} and products $U = U_1 \cdots U_t$ for the U_i in $R(\mathfrak{A})$ and $L(\mathfrak{A})$. Then $EU = 0$. Hence every transformation of $T(\mathfrak{A})$ is expressible as a sum $S + \alpha I$ with $ES = 0$ and α in \mathfrak{F} . Since $T(\mathfrak{A})$ contains $I\mathfrak{F}$ we have S in $T(\mathfrak{A})$, and (18) holds. That \mathfrak{S} is an ideal of \mathfrak{A} follows from the property that if U and S are in \mathfrak{S} then $E[U(S + \alpha I)] = EU(S + \alpha I) = 0$, $E[(S + \alpha I)U] = ESU + EU\alpha = 0$.

7. Simple algebras

An algebra \mathfrak{A} is said to be *simple* if \mathfrak{A} is not a zero algebra of order one and \mathfrak{A} is the only non-zero ideal of \mathfrak{A} . We define the \mathfrak{A} -centralizer of any set \mathfrak{S} of

quantities of any algebra \mathfrak{A} to be the set of all quantities k in \mathfrak{A} such that $k \cdot h = h \cdot k$ for every h of \mathfrak{F} and see that this set is a subalgebra of \mathfrak{A} if \mathfrak{A} is associative. Then we may prove

LEMMA 10. *The algebra $T(\mathfrak{A})$ is simple if and only if \mathfrak{A} is either simple or $T(\mathfrak{A}) = I\mathfrak{F}$ and \mathfrak{A} is a zero algebra.*

For let $T(\mathfrak{A})$ be simple. By Lemma 9 if \mathfrak{A} contains an absolute divisor of zero we have $\mathfrak{S} = 0$ in (18) and \mathfrak{A} is a zero algebra, by Lemma 8. Hence let \mathfrak{A} be not a zero algebra and suppose that $\mathfrak{B} = \mathfrak{A}E$ is a non-zero ideal of \mathfrak{A} for an idempotent $E \neq 0$ of $(\mathfrak{F})_n$. We define \mathfrak{S} to be the set of all S in $T(\mathfrak{A})$ such that $S = \mathfrak{S}E$. By (17) \mathfrak{S} contains $L(\mathfrak{B}, \mathfrak{A})$ and similarly contains $R(\mathfrak{B}, \mathfrak{A})$. But if $y \neq 0$ and $L_y = 0$ we have $R_y \neq 0$ since \mathfrak{A} contains no absolute divisor of zero. Hence $L(\mathfrak{B}, \mathfrak{A})$ and $R(\mathfrak{B}, \mathfrak{A})$ are not both zero, $\mathfrak{S} \neq 0$. If S is in \mathfrak{S} and U is in $T(\mathfrak{A})$ we have $SU = SEU = SEUE = (SU)E$ by Lemma 4 while also $US = (US)E$. But then SU and US are in \mathfrak{S} , \mathfrak{S} is an ideal of $T(\mathfrak{A})$, $\mathfrak{S} = T(\mathfrak{A})$ contains $I = IE = E$, $\mathfrak{B} = \mathfrak{A}E = \mathfrak{A}$ is simple.

Conversely let \mathfrak{A} be simple. If $T(\mathfrak{A})$ has a nilpotent ideal \mathfrak{N} we have $\mathfrak{NR}(\mathfrak{A})$ in \mathfrak{N} , $\mathfrak{NL}(\mathfrak{A})$ in \mathfrak{N} so that if $\mathfrak{B} = \mathfrak{AN}$ we have $\mathfrak{BA} = \mathfrak{BR}(\mathfrak{A}) = \mathfrak{ANR}(\mathfrak{A})$ contained in \mathfrak{B} , $\mathfrak{AB} = \mathfrak{BL}(\mathfrak{A}) = \mathfrak{ANL}(\mathfrak{A})$ in \mathfrak{B} . Hence \mathfrak{B} is an ideal of \mathfrak{A} . But by Lemma 1 $\mathfrak{B} \neq 0$, \mathfrak{A} contrary to hypothesis. Hence $T(\mathfrak{A})$ is an associative semi-simple algebra and is either simple as desired or is a direct sum. In the latter case the unity quantity of a component of $T(\mathfrak{A})$ is an idempotent E , in the $(\mathfrak{F})_n$ -centralizer of $T(\mathfrak{A})$, which is singular and not zero. Then by Lemma 4 $\mathfrak{B} = \mathfrak{A}E$ is an ideal of \mathfrak{A} , $\mathfrak{B} \neq 0$ or \mathfrak{A} . This completes our proof.

8. Central simple algebras

A field \mathfrak{C} consisting of linear transformations over \mathfrak{F} on a linear space \mathfrak{A} of order n over \mathfrak{F} is called a *subfield* of $(\mathfrak{F})_n$ if the identity transformation of $(\mathfrak{F})_n$ is in \mathfrak{C} . Then \mathfrak{A} may be regarded as being a linear space of order σ over \mathfrak{C} and $n = \sigma\tau$ where τ is the degree of \mathfrak{C} over \mathfrak{F} . The set of all linear transformations over \mathfrak{C} on \mathfrak{A} is the total matrix algebra $(\mathfrak{C})_\sigma$ and is clearly the $(\mathfrak{F})_n$ -centralizer of \mathfrak{C} . The equations $a \cdot x = aR_x$, $x \cdot a = aL_x$ then define the algebra \mathfrak{A} over \mathfrak{F} as an algebra over \mathfrak{C} if and only if every R_x and L_x is in $(\mathfrak{C})_\sigma$. But then $R(\mathfrak{A})$, $L(\mathfrak{A})$, $T(\mathfrak{A})$ are in $(\mathfrak{C})_\sigma$. It follows that \mathfrak{A} is an algebra over a subfield \mathfrak{C} of $(\mathfrak{F})_n$ if and only if \mathfrak{C} is in the $(\mathfrak{F})_n$ -centralizer of $T(\mathfrak{A})$.

We define the *transformation center* of \mathfrak{A} to be the $T(\mathfrak{A})$ -centralizer of $T(\mathfrak{A})$ and designate this subalgebra of $T(\mathfrak{A})$ by $\mathfrak{C}(\mathfrak{A})$. If $T(\mathfrak{A})$ is simple the transformation center of \mathfrak{A} is a field of degree t over \mathfrak{F} and $n = st$, $T(\mathfrak{A})$ is contained in the total matrix algebra $[\mathfrak{C}(\mathfrak{A})]_s$.

An algebra \mathfrak{A} over \mathfrak{F} is said to be *central simple over \mathfrak{F}* if $\mathfrak{A}_\mathfrak{K}$ is simple for every scalar extension \mathfrak{K} of \mathfrak{F} . If then \mathfrak{A} is simple and \mathfrak{B} is any subfield of its transformation center $\mathfrak{C} = \mathfrak{C}(\mathfrak{A})$ the degree of \mathfrak{B} divides $t = \tau\rho$ and \mathfrak{A} is simple of order $s\rho$ over \mathfrak{B} , $T(\mathfrak{A})$ is simple over \mathfrak{B} and is contained in $\mathfrak{B}_{s\rho}$. But $[T(\mathfrak{A})]_\mathfrak{K} = T(\mathfrak{A}_\mathfrak{K})$ for every scalar extension \mathfrak{K} of \mathfrak{B} and thus \mathfrak{A} is central simple over \mathfrak{B}

only if $T(\mathfrak{A})$ is central simple over \mathfrak{Z} . This can occur only if $\mathfrak{Z} = \mathfrak{C}(\mathfrak{A})$. We use this result and then prove⁶

THEOREM 1. *An algebra \mathfrak{A} of order $n > 1$ over \mathfrak{F} is simple if and only if $T(\mathfrak{A})$ is the total matrix algebra $(\mathfrak{C})_s$ where $\mathfrak{C} = \mathfrak{C}(\mathfrak{A})$ is a field of degree t over \mathfrak{F} and $n = st$. Moreover \mathfrak{A} is central simple over a subfield \mathfrak{Z} of $(\mathfrak{F})_n$ if and only if $\mathfrak{Z} = \mathfrak{C}$.*

For if \mathfrak{A} is simple so is \mathfrak{A} over its transformation center \mathfrak{C} and $\mathfrak{A}_{\mathfrak{R}}$ is not a zero algebra for any scalar extension \mathfrak{R} of \mathfrak{C} . Now $T(\mathfrak{A})$ is in $(\mathfrak{C})_s$ and is known⁷ to be a central simple algebra over \mathfrak{C} . The $(\mathfrak{C})_s$ -centralizer of $T(\mathfrak{A})$ is also a central simple algebra \mathfrak{D} of degree q over \mathfrak{C} and $T(\mathfrak{A}) = (\mathfrak{C})_s$ if and only if $q = 1$. If $q > 1$ we let \mathfrak{R} be a splitting field over \mathfrak{C} of \mathfrak{D} and see that the total matrix algebra $\mathfrak{D}_{\mathfrak{R}}$ contains a non-zero idempotent E which is singular and in the $(\mathfrak{C})_s$ -centralizer of $T(\mathfrak{A}_{\mathfrak{R}})$. Then by Lemma 4 $\mathfrak{A}_{\mathfrak{R}}E$ is a non-zero proper ideal over \mathfrak{R} of $\mathfrak{A}_{\mathfrak{R}}$, whereas the proof above shows that $\mathfrak{A}_{\mathfrak{R}}$ is simple, a contradiction. It follows that $T(\mathfrak{A}) = (\mathfrak{C})_s$. The only subfields \mathfrak{Z} of $(\mathfrak{F})_n$ in the $(\mathfrak{F})_n$ -centralizer of $T(\mathfrak{A})$ are in the $(\mathfrak{F})_n$ -centralizer of \mathfrak{C} and hence in $(\mathfrak{C})_s$. They are then in the $(\mathfrak{C})_s$ -centralizer of $(\mathfrak{C})_s$ and thus in \mathfrak{C} , \mathfrak{A} is central simple over \mathfrak{Z} only if $\mathfrak{Z} = \mathfrak{C}$. Conversely let $T(\mathfrak{A}) = (\mathfrak{C})_s$ where $n = st$ and $(\mathfrak{C})_s$ is a total matrix algebra of degree s over \mathfrak{C} , \mathfrak{C} is a field of degree t over \mathfrak{F} . Then the order of $T(\mathfrak{A})$ is $s^2t > 1$ since otherwise $s = t = n = 1$. Hence \mathfrak{A} is not a zero algebra and, by Lemma 9, \mathfrak{A} is simple. Also \mathfrak{A} is central simple over \mathfrak{C} since $T(\mathfrak{A})$ is a total matrix algebra over \mathfrak{C} and is central simple, $\mathfrak{A}_{\mathfrak{R}}$ is not a zero algebra over any scalar extension \mathfrak{R} of \mathfrak{C} , $\mathfrak{A}_{\mathfrak{R}}$ is simple. Thus \mathfrak{A} is central simple over its transformation center.

9. Algebras with a left unity quantity

Let e be a non-zero vector in a linear space of order n over \mathfrak{F} and \mathfrak{S} be any linear subspace of order $m \leq n$ of $(\mathfrak{F})_n$. Then $e\mathfrak{S}$ is a linear subspace of \mathfrak{L} and the correspondence

$$(19) \quad S \rightarrow eS \quad (S \text{ in } \mathfrak{S}),$$

is a linear mapping of \mathfrak{S} on $e\mathfrak{S}$. It follows that the order of $e\mathfrak{S}$ over \mathfrak{F} is at most m .

⁶ Results essentially equivalent to this one and to Lemma 10 were given by N. Jacobson, *A note on non-associative algebras*, Duke J., vol. 3 (1937), pp. 544-8. The result was first announced for Lie algebras by the author in the A. M. S. Bulletin, vol. 41 (1935), p. 344. and the author feels that the present exposition is not only in a form better suited than that of Jacobson for later application but presents also a much clearer picture of the relations between an algebra \mathfrak{A} and its transformation algebra $T(\mathfrak{A})$. The idea of studying these relations was suggested to both Jacobson and the author by the lectures of H. Weyl on Lie Algebras which were given in Fine Hall in 1933. Note that if \mathfrak{I} is the algebra generated by the right and left multiplications of \mathfrak{A} then $\mathfrak{I} = T(\mathfrak{A})$ unless \mathfrak{I} does not contain the identity transformation. But then \mathfrak{I} is an ideal of $T(\mathfrak{A})$ and this cannot occur when \mathfrak{A} is simple.

⁷ For the properties used here see Chapters I, III, IV of the author's *Structure of Algebras*.

Suppose then that (19) is one-to-one, and that $m = n$. Then (19) maps \mathfrak{S} on \mathfrak{I} such that $x = eS = eT$ if and only if $S = T$. We may then define R_x as that transformation S for which $eS = x$ and have defined an algebra \mathfrak{A} without absolute right divisors of zero and such that $\mathfrak{S} = R(\mathfrak{A})$, $e \cdot x = eR_x = x$ for every x of \mathfrak{A} .

The quantity e is now a left unity quantity of \mathfrak{A} . Conversely every algebra \mathfrak{A} with a left unity quantity e has no absolute right divisors of zero and is such that the linear mapping $R_x \rightarrow x = eR_x$ is a one-to-one mapping of $R(\mathfrak{A})$ on $\mathfrak{A} = eR(\mathfrak{A})$. Then multiplication in \mathfrak{A} is given by

$$(20) \quad eS \cdot eR = eSR$$

for every S of $(\mathfrak{F})_n$ and every R of $R(\mathfrak{A})$. It may be seen, however, that (20) does not hold if R is not in $R(\mathfrak{A})$.

If e is given we say that b in \mathfrak{A} is *right regular with respect to e* if $c \cdot b = e$ for c in \mathfrak{A} . Then c is a *left inverse* of b (relative to e). Such a quantity may exist even when b is *right singular*.

The left inverse of a right non-singular quantity b may be expressed as a certain polynomial in b . We define $b^2 = bb$ and then the right powers of b by $b^{k+1} = (b^k)b$ for $k > 0$, $b^k = e$ for $k = 0$. If λ is an indeterminate over \mathfrak{F} and

$$(21) \quad \phi(\lambda) = \lambda^t + \beta_1 \lambda^{t-1} + \cdots + \beta_t \quad (\beta_i \text{ in } \mathfrak{F})$$

we define the right polynomial

$$(22) \quad \phi_R(b) = b^t + \beta_1 b^{t-1} + \cdots + \beta_{t-1} b + \beta_t e$$

for any b in \mathfrak{A} and right powers b^k . Then it is clear from (26) that

$$(23) \quad \phi_R(b) = e\phi(R_b).$$

Hence $\phi_R(b) = 0$ if $\phi(R_b) = 0$. However $\phi_R(b)$ may be zero for $\phi(R_b)$ a non-zero linear transformation carrying the vector e into zero. Note now that in particular

$$(24) \quad f_R(b) = g_R(b) = 0$$

where $f(\lambda)$ is the characteristic function and $g(\lambda)$ is the minimum function of R_b .

We define the *right minimum function* (with respect to e) of a quantity b of \mathfrak{A} to be the polynomial (21) of least degree t such that $\phi_R(b) = 0$. Its uniqueness is then implied by

THEOREM 2. *The right minimum function $\phi(\lambda)$ of a quantity b of an algebra \mathfrak{A} divides every $\psi(\lambda)$ such that $\psi_R(b) = 0$.*

For we write $\psi(\lambda) = \phi(\lambda)\rho(\lambda) + \sigma(\lambda)$ and have $\psi_R(b) = e\psi(R_b) = e[\phi(R_b)\rho(R_b) + \sigma(R_b)] = e\sigma(R_b) = \sigma_R(b)$. But the degree of $\sigma(\lambda)$ may be taken to be less than that of $\phi(\lambda)$, $\sigma(\lambda)$ is identically zero.

We see in particular that the right minimum function $\phi(\lambda)$ of b divides the minimum and characteristic functions of R_b . If b is right non-singular the

constant term of these latter functions is not zero and $\phi(\lambda)$ has the form (27) for $\beta_i = 0$. But then $\phi_R(b) = (b^{t-1} + \beta_1 b^{t-2} + \cdots + \beta_{t-1} e) \cdot b + \beta_t e = 0$,

$$(25) \quad b^{-1} \cdot b = e, \quad b^{-1} = -\beta_t^{-1}(b^{t-1} + \beta_1 b^{t-2} + \cdots + \beta_{t-1} e).$$

Moreover if $c \cdot b = e$ we have $(c - b^{-1}) \cdot b = (c - b^{-1})R_b = 0$ if and only if $c = b^{-1}$.

Right singular quantities may be right regular and it may happen that, while (25) holds for $\beta_i \neq 0$, R_b may be singular. To illustrate this we consider the linear space \mathfrak{L} of order three over a field \mathfrak{F} of characteristic not two where \mathfrak{L} consists of all two-rowed symmetric matrices. Then a basis of \mathfrak{L} is given by

$$(26) \quad e = u_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad u_2 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad u_3 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

We define an algebra \mathfrak{A} by

$$(27) \quad a \cdot x = \frac{ax + xa}{2}$$

where products on the right are ordinary two-rowed square matrix products. Then \mathfrak{A} is a commutative algebra such that $e \cdot x = x \cdot e = \frac{1}{2}(ex + xe) = x$. But also $x \cdot x = x^2$ and thus

$$(28) \quad u_2 \cdot u_2 = u_3 \cdot u_3 = e,$$

while

$$(29) \quad 2u_2 \cdot u_3 = 2u_3 \cdot u_2 = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = 0.$$

It follows that u_2 and u_3 are right and left divisors of zero and are right (and left) regular. The (right) minimum function of both u_2 and u_3 is $\lambda^2 - 1$ and the minimum function of R_{u_2} and R_{u_3} is $\lambda(\lambda^2 - 1)$. Here the matrices of these linear transformations with respect to the basis (32) are

$$\Gamma_{u_2} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \Gamma_{u_3} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$

10. Algebras with a unity quantity

A quantity f of \mathfrak{A} is a *right unity quantity* of \mathfrak{A} if $x \cdot f = x$ for every x of \mathfrak{A} , that is, $R_f = I$. Thus the mapping

$$(30) \quad L_x \mapsto fL_x = x$$

of $L(\mathfrak{A})$ on \mathfrak{A} is one-to-one, and multiplication in \mathfrak{A} is defined by

$$(31) \quad eL \cdot eS = eSL$$

for every S of $(\mathfrak{F})_n$ and every L of $R(\mathfrak{A})$. *Left regularity* and *left polynomials* are defined in the obvious fashion and every left non-singular quantity b has a right inverse which is a left polynomial in b .

If \mathfrak{A} has both a left unity quantity e and a right unity quantity f we have both $e \cdot f = f$ and $e \cdot f = e$ so that $e = f$. Then e is the unique quantity of \mathfrak{A} such that $e \cdot x = x \cdot e$ for every x of \mathfrak{A} and we shall call e the *unity quantity* of \mathfrak{A} . It has the determining property

$$(32) \quad R_e = L_e = I,$$

and multiplication in \mathfrak{A} is defined by

$$(33) \quad eS \cdot eR = eSR, \quad eL \cdot eS = eSL$$

for every S of $(\mathfrak{F})_n$, R of $R(\mathfrak{A})$, L of $L(\mathfrak{A})$. Note that then

$$(34) \quad eL \cdot eR = eRL = eLR,$$

so that $e(RL - LR) = 0$ for every R of $R(\mathfrak{A})$ and L of $L(\mathfrak{A})$. However we shall see that $RL = LR$ for every such R and L if and only if \mathfrak{A} is associative.

11. Isotopes of algebras

All algebras \mathfrak{A} of the same order n may be regarded as having quantities comprising the same linear space L of order n over \mathfrak{F} . If \mathfrak{A} is given the space $R(\mathfrak{A})$ and the mapping $x \rightarrow R_x$ of L on $R(\mathfrak{A})$ are thereby determined and conversely. Thus if \mathfrak{A}_0 is a second algebra we have a corresponding linear mapping $x \rightarrow R_x^{(0)}$ of \mathfrak{A}_0 on a linear subspace $R(\mathfrak{A}_0)$ of $(\mathfrak{F})_n$ and we write

$$(35) \quad (a, x) = aR_x^{(0)}$$

for products in \mathfrak{A}_0 . We shall now say that \mathfrak{A} is *isotopic* to \mathfrak{A}_0 if there exist non-singular linear transformations P, Q, C such that

$$(36) \quad R_x^{(0)} = PR_xQC,$$

and shall call (36) an *isotopy* of \mathfrak{A} and \mathfrak{A}_0 .

If \mathfrak{A} is isotopic to \mathfrak{A}_0 then $R_{xQ^{-1}}^{(0)} = PR_xC$ so that $R_x = P^{-1}R_{xQ^{-1}}C^{-1}$ and \mathfrak{A}_0 is isotopic to \mathfrak{A} . Also \mathfrak{A} is isotopic to itself under (36) with $P = Q = C = I$. Finally if $R_x^{(1)} = P_1R_{xQ_1}^{(0)}C_1$ then $R_x^{(1)} = P_2R_{xQ_2}^{(0)}C_2$ where $P_2 = P_1P$, $Q_2 = Q_1Q$, $C_2 = CC_1$. Hence the relation of isotopy is a formal equivalence relation and we shall say that \mathfrak{A} and \mathfrak{A}_0 are isotopic as well as that \mathfrak{A} is isotopic to \mathfrak{A}_0 .

All left multiplications $L_x^{(0)}$ of \mathfrak{A}_0 are determined when its right multiplications are given and conversely. Thus we shall determine the conditions relating $L_x^{(0)}$ and L_x which are equivalent to (36). We observe that in \mathfrak{A} we have

$$(37) \quad a \cdot x = aR_x = xL_a, \quad x \cdot a = aL_x = xR_a,$$

and in \mathfrak{A}_0 we have

$$(38) \quad (a, x) = aR_x^{(0)} = xL_a^{(0)}, \quad (x, a) = aL_x^{(0)} = xR_a^{(0)}.$$

Define

$$(39) \quad b = aQ, \quad z = xP$$

and obtain

$$(40) \quad aL_x^{(0)} = xR_a^{(0)} = xPR_bC = zR_bC = (z \cdot b)C = bL_zC.$$

Then $aL_x^{(0)} = aQL_zC$ for every a and x of \mathfrak{A} and we have

THEOREM 3. *The conditions (36) which imply that \mathfrak{A} and \mathfrak{A}_0 are isotopic are equivalent to*

$$(41) \quad L_x^{(0)} = QL_zPC.$$

The relation of equivalence is an instance of isotopy. For two algebras \mathfrak{A}_0 and \mathfrak{A} are said to be equivalent if there exists a non-singular linear transformation $a \rightarrow aH$ on \mathfrak{A}_0 to \mathfrak{A} which is preserved under multiplication. But then

$$(42) \quad (a, x)H = aH \cdot xH,$$

that is $aR_x^{(0)}H = aHR_{xH}$. Hence \mathfrak{A}_0 and \mathfrak{A} are equivalent if and only if

$$(43) \quad R_x^{(0)} = HR_{xH}H^{-1}.$$

By Theorem 3 the equivalent algebras \mathfrak{A}_0 and \mathfrak{A} are also related by

$$(44) \quad L_x^{(0)} = HL_{xH}H^{-1}.$$

It is usually more convenient to use simplifications of (36) and (41) obtainable by replacing \mathfrak{A}_0 by an equivalent algebra. Thus we may apply (43) to (36) with $H = Q^{-1}$ and have $R_x^{(1)} = HR_{xH}^{(0)}H^{-1} = (HP)R_xCH^{-1}$. This result together with Theorem 3 may be stated as

THEOREM 4. *Every isotope of an algebra \mathfrak{A} is equivalent to an isotope defined by*

$$(45) \quad R_x^{(0)} = PR_xC, \quad L_x^{(0)} = L_{xP}C,$$

for non-singular linear transformations P and C .

The form above⁸ has the advantage that in \mathfrak{A}_0 we map x on PR_xC but is so unsymmetrical that we shall prefer the *principal isotopy* obtained from (36) by the application of (43) with $H = C$. We state the result as

⁸ The concept of isotopy was suggested to the author by the work of N. Steenrod who, in his study of homotopy groups in topology, was led to study isotopy of division algebras. He concluded that algebras related as in (45) would yield the same homotopy properties and should therefore be put in the same class. The author then formulated the concept generally as in (36) and obtained Theorem 3 giving the corresponding property for left multiplications and Theorem 4 showing that Steenrod's isotopes were actually equivalent to the more general type. However the principal isotopes of (46) are much more conveniently handled.

THEOREM 5. *Every isotope of an algebra \mathfrak{A} is equivalent to a principal isotope \mathfrak{A}_0 , that is, an isotope with*

$$(46) \quad R_x^{(0)} = PR_xQ, \quad L_x^{(0)} = QL_xP,$$

for non-singular linear transformations P and Q .

We observe that (46) implies that $R_x = P^{-1}R_{xQ}^{(0)}$, $L_x = Q^{-1}L_{xP}^{(0)}$ and thus that if \mathfrak{A}_0 is a principal isotope of \mathfrak{A} then \mathfrak{A} is a principal isotope of \mathfrak{A}_0 . If also $R_x^{(1)} = UR_{xV}^{(0)}$ we have $R_x^{(1)} = (UP)R_{xVQ}$, $L_x^{(1)} = VL_{xU}^{(0)} = VQL_{xUP}$. Finally \mathfrak{A} is a principal isotope of itself with P and Q the identity. Thus we shall again say that \mathfrak{A} and \mathfrak{A}_0 are principal isotopes as well as that \mathfrak{A}_0 is a principal isotope of \mathfrak{A} .

Let us note in closing this section that every automorphism of an algebra \mathfrak{A} is an equivalence H of \mathfrak{A} and itself. Then $(a, x) = a \cdot x$ and $R_x^{(0)} = R_x$, $L_x^{(0)} = L_x$. We state this result as

THEOREM 6. *A linear transformation H on an algebra \mathfrak{A} defines an automorphism of \mathfrak{A} if and only if H is non-singular and such that either (and hence both) of the following conditions holds*

$$(47) \quad R_{xH} = H^{-1}R_xH, \quad L_{xH} = H^{-1}L_xH \quad (x \text{ in } \mathfrak{A}).$$

12. Isotopes with a unity quantity

If \mathfrak{A} has a unity quantity e and f is any non-zero quantity of \mathfrak{A} there exists a non-singular linear transformation H such that $e = fH$. Then (43) and (44) imply that $R_f^{(0)} = HR_eH^{-1} = L_f^{(0)} = HL_eH^{-1} = I$ since $R_e = L_e = I$. It follows that \mathfrak{A} is equivalent to an algebra \mathfrak{A}_0 with f as unity quantity. However we seek to discover what principal isotopes of \mathfrak{A} have f as unity quantity. We shall obtain the answer to this question in

THEOREM 7. *Let g range over all left non-singular quantities of \mathfrak{A} , h range over all right non-singular quantities of \mathfrak{A} , so that the non-singular linear transformations*

$$(48) \quad P = (R_h)^{-1}, \quad Q = (L_g)^{-1}$$

exist for each g and h . Then the principal isotope of \mathfrak{A} defined by (46), (48) has $f = g \cdot h$ as a unity quantity. Conversely every isotope of \mathfrak{A} with a unity quantity f is equivalent to a principal isotope determined as in (48), (46) for $f = g \cdot h$.

For if $f = g \cdot h$ we have $f = gR_h = hL_g$ and $g = fP$, $h = fQ$ where P and Q are defined in (48). Let \mathfrak{A}_0 be the principal isotope of \mathfrak{A} defined by (46) for this P and Q and put $x = f$. Then

$$R_f^{(0)} = PR_h = L_f^{(0)} = QL_g = I,$$

f is the unity quantity of \mathfrak{A}_0 . Conversely let (46) define an isotope of \mathfrak{A} with f as its unity quantity so that if we define $g = fP$, $h = fQ$ we have $R_f^{(0)} = I = PR_h$, $L_f^{(0)} = I = QL_g$. But then h is right non-singular, g is left non-singular, (54) holds and $f = gP^{-1} = gR_h = g \cdot h$ as desired.

We now prove

THEOREM 8. *Let \mathfrak{A} and \mathfrak{A}_0 be principal isotopes and let each of these algebras have a unity quantity. Then the corresponding transformation algebras $T(\mathfrak{A})$ and $T(\mathfrak{A}_0)$ are the same.*

For we have (48) and hence have P and Q in $T(\mathfrak{A})$, $R_x^{(0)}$ and $L_x^{(0)}$ in $T(\mathfrak{A})$, $T(\mathfrak{A}_0)$ is contained in $T(\mathfrak{A})$. The converse follows by symmetry.

13. Ideals in isotopes

The mapping $x \rightarrow R_x$ of \mathfrak{A} on $R(\mathfrak{A})$ is one-to-one if and only if $x \rightarrow R_x^{(0)} = PR_xQ$ is one-to-one. Hence \mathfrak{A} contains no absolute right divisors of zero if and only if every isotope of \mathfrak{A} has this property. In particular \mathfrak{A} is a zero algebra if and only if every isotope of \mathfrak{A} is a zero algebra. We combine this result with those of Theorems 1, 5, 8 to obtain

THEOREM 9. *Let \mathfrak{A} and \mathfrak{A}_0 be isotopic algebras each possessing a unity quantity. Then \mathfrak{A} is simple if and only if \mathfrak{A}_0 is simple.*

We also have the stronger result

THEOREM 10. *Let \mathfrak{A} and \mathfrak{A}_0 be principal isotopes and let each have a unity quantity. Then a linear subspace of \mathfrak{A} is an ideal of \mathfrak{A} if and only if it is an ideal of \mathfrak{A}_0 .*

This follows from Lemma 4 and from Theorem 8. That it is desirable whenever possible to restrict our attention to algebras with a unity quantity is strongly indicated by the remarkable

THEOREM 11. *Let there exist a polynomial $f(x)$ of degree n over \mathfrak{F} which is irreducible in \mathfrak{F} . Then every algebra \mathfrak{A} of order n over \mathfrak{F} and with a unity quantity e has a principal isotope which is simple and indeed has neither left nor right ideals.*

For by Lemma 6 if we take the linear transformation P of $(\mathfrak{F})_n$ such that $f(P) = 0$ the only idempotents E such that $EP = EPE$ are 0, I . Define \mathfrak{A}_0 by (46) for this P and $Q = P$. Then $R_{eP^{-1}}^{(0)} = PR_e = P$, $L_{eP^{-1}}^{(0)} = QL_e = P$ and $ER(\mathfrak{A}_0) = ER(\mathfrak{A}_0)E$ is not possible unless $EP = EPE$, $EL(\mathfrak{A}_0) = EL(\mathfrak{A}_0)E$ is not possible unless $EP = EPE$. Hence in either case $E = 0, I$, the only right and left ideals of \mathfrak{A}_0 are zero and \mathfrak{A}_0 .

14. Associative algebras

It is well known⁹ that if \mathfrak{A} is an associative algebra with a unity quantity the space $R(\mathfrak{A})$ is an algebra and the mapping $x \rightarrow R_x$ defines an equivalence of \mathfrak{A} and $R(\mathfrak{A})$. Moreover $L(\mathfrak{A})$ is also an algebra and $x \rightarrow L_x$ defines a reciprocal simple isomorphism of \mathfrak{A} and $L(\mathfrak{A})$. However it is possible for $R(\mathfrak{A})$ to be an algebra without \mathfrak{A} being associative. We shall give an illustration of such an occurrence shortly.

Let us now observe the known criterion for associativity which we state as

LEMMA 11. *Let $R(\mathfrak{A})$ and $L(\mathfrak{A})$ be the right and left multiplication spaces*

⁹ Cf. the reference in footnote 7.

respectively of an algebra \mathfrak{A} . Then \mathfrak{A} is associative if and only if $RL = LR$ for every R of $R(\mathfrak{A})$ and L of $L(\mathfrak{A})$.

The proof of this criterion is rather immediate. We write $(x \cdot a) \cdot y = x \cdot (a \cdot y)$ for every a, x, y of \mathfrak{A} and see that this equation is equivalent to

$$(49) \quad (aL_x)R_y = x(aR_y) = aR_yL_x.$$

Thus $L_xR_y = R_yL_x$ as desired.

We now derive the important

THEOREM 12. *An algebra \mathfrak{A} with a unity quantity is associative if and only if every isotope with a unity quantity of \mathfrak{A} is associative and equivalent to \mathfrak{A} .*

For if \mathfrak{A} is associative $R_xR_y = R_{xy}$, $L_xL_y = L_{yx}$ for every x and y of \mathfrak{A} . A quantity x of \mathfrak{A} is right non-singular if and only if it has an inverse in \mathfrak{A} and then x is also left non-singular. But then

$$(50) \quad R_{x^{-1}} = (R_x)^{-1}, \quad L_{x^{-1}} = (L_x)^{-1}.$$

Let now \mathfrak{A}_0 be a principal isotope of \mathfrak{A} and assume that \mathfrak{A}_0 has a unity quantity so that (48) holds. Then $P = R_{h^{-1}}$, $Q = L_{g^{-1}}$ and we have $xQ = xL_{g^{-1}} = g^{-1} \cdot x$, $PR_{xQ} = R_{h^{-1}}R_{xQ} = R_{h^{-1} \cdot g^{-1} \cdot x}$. However $f^{-1} = (g \cdot h)^{-1} = h^{-1} \cdot g^{-1}$ and we have proved that

$$(51) \quad R_x^{(0)} = R_{f^{-1}x}.$$

Similarly $L_x^{(0)} = L_{g^{-1}L_{xP}} = L_{g^{-1}L_{x \cdot h^{-1}}} = L_{x \cdot h^{-1} \cdot g^{-1}}$,

$$(52) \quad L_x^{(0)} = L_{x \cdot f^{-1}}.$$

It follows that $R(\mathfrak{A}_0) = R(\mathfrak{A})$, $L(\mathfrak{A}_0) = L(\mathfrak{A})$, $R_x^{(0)}L_y^{(0)} = L_y^{(0)}R_x^{(0)}$ for every x and y of \mathfrak{A} . By Lemma 11 the algebra \mathfrak{A}_0 is associative. Since it has a unity quantity it is equivalent to the algebra $R(\mathfrak{A}_0) = R(\mathfrak{A})$ which is equivalent to \mathfrak{A} , \mathfrak{A} and \mathfrak{A}_0 are equivalent.

Observe that if $H = R_{f^{-1}}$ then $xH = x \cdot f^{-1}$, $HR_{xH}H^{-1} = R_{f^{-1}}R_{x \cdot f^{-1}}R_f = R_{f^{-1} \cdot x}$. Hence

$$(53) \quad R_x^{(0)} = HR_{xH}H^{-1}$$

and the principal isotopy of \mathfrak{A} and \mathfrak{A}_0 which we are studying is induced by the linear mapping

$$x \rightarrow xH = xf^{-1}$$

of \mathfrak{A}_0 on \mathfrak{A} . This map is an equivalence of \mathfrak{A}_0 and \mathfrak{A} obtainable as the product of the equivalence $x \rightarrow R_x^{(0)}$ of \mathfrak{A}_0 on $R(\mathfrak{A}_0) = R(\mathfrak{A})$, the automorphism $R_x^{(0)} = HR_{xH}H^{-1} \rightarrow R_{xH}$ of $R(\mathfrak{A})$ and the equivalence $R_{xH} \rightarrow xH$ of $R(\mathfrak{A})$ on \mathfrak{A} . Observe also that the only principal isotopy of an associative algebra with a unity quantity e which carries e into the unity quantity of the isotope is that given by $R_x^{(0)} = R_x$. For (51) holds with $f = e$, $f^{-1} \cdot x = e \cdot x = x$.

It is natural to consider at this point whether the property that $R(\mathfrak{A})$ is an algebra implies that \mathfrak{A} is an associative algebra. This is partially true in view of

THEOREM 13. *An algebra \mathfrak{A} with a left unity quantity e is associative if and only if $R(\mathfrak{A})$ is an algebra. Moreover the mapping $x \rightarrow R_x$ then is an equivalence of \mathfrak{A} and $R(\mathfrak{A})$.*

For we have $e \cdot x = eR_x = x$, the mapping $R_x \rightarrow eR_x = x$ is a one-to-one linear mapping of \mathfrak{A} on $R(\mathfrak{A})$. Now $(e \cdot x) \cdot y = x \cdot y = eR_x R_y = e \cdot (x \cdot y) = eR_{x \cdot y}$ and $R_{x \cdot y}$ is in $R(\mathfrak{A})$. It follows that $R_{x \cdot y} = R_x R_y$, \mathfrak{A} is equivalent to $R(\mathfrak{A})$ and is associative. The converse has already been mentioned.

We may also prove the simple generalization

THEOREM 14. *Let \mathfrak{A} and $R(\mathfrak{A})$ be algebras and let there be a left non-singular quantity f in \mathfrak{A} . Then \mathfrak{A} has an associative principal isotope \mathfrak{A}_0 which is equivalent to $R(\mathfrak{A})$ and has f as left unity quantity.*

For we define \mathfrak{A}_0 by $R_x^{(0)} = R_x Q$, $Q = (L_f)^{-1}$. Then $L_x^{(0)} = QL_x = (L_f)^{-1}L_x$ and hence $L_f^{(0)} = I$, $(f, x) = xL_f^{(0)} = x$, \mathfrak{A}_0 has f as its left unity quantity. But $R(\mathfrak{A}) = R(\mathfrak{A}_0)$ and our result follows from Theorem 13.

It remains to consider the general question as to the existence of non-associative algebras \mathfrak{A} such that $R(\mathfrak{A})$ is an algebra. Such algebras do exist and we prove this as an immediate consequence of

THEOREM 15. *Let \mathfrak{A} be an associative algebra of order $n > 1$ over an infinite field \mathfrak{F} and let \mathfrak{A} have a unity quantity e so that $R(\mathfrak{A})$ is an algebra. Then there exists a non-associative isotope \mathfrak{A}_0 of \mathfrak{A} with $R(\mathfrak{A}_0) = R(\mathfrak{A})$.*

For it is known⁹ that $L(\mathfrak{A})$ is the $(\mathfrak{F})_n$ -centralizer of $R(\mathfrak{A})$, $L(\mathfrak{A})$ is a proper subalgebra of $(\mathfrak{F})_n$. Then there exists a linear transformation U not in $L(\mathfrak{A})$, $UR_a - R_a U \neq 0$ for some a in \mathfrak{A} . Every linear transformation has the form

$$U = \sum_{i=1}^{n^2} \xi_i S_i \quad (\xi_i \text{ in } \mathfrak{F})$$

for a basis S_i of $(\mathfrak{F})_n$, and $UR_a - R_a U \neq 0$ implies that there exist η_i in \mathfrak{F} such that $Q = \sum \eta_i S_i$ is non-singular, $QR_a \neq R_a Q$. We define \mathfrak{A}_0 by $R_x^{(0)} = R_x Q$ and have $R(\mathfrak{A}_0) = R(\mathfrak{A})$, $L_x^{(0)} = QL_x$, $L_e^{(0)} = Q$ is not commutative with $R_a^{(0)-1}$, \mathfrak{A}_0 is not associative.

It is important to observe that *there exist associative algebras (without unity quantities) which are isotopic but not equivalent.* For example consider the nilpotent algebra \mathfrak{A} with basis e_1, e_2, e_3 such that $e_1 \cdot e_2 = -e_2 \cdot e_1 = e_3$ and all other products are zero. Then we write $a = (\alpha_1, \alpha_2, \alpha_3)$ for $a = \alpha_1 e_1 + \alpha_2 e_2 + \alpha_3 e_3$ and have $a \cdot x = a \cdot (\xi_1, \xi_2, \xi_3) = (0, 0, \alpha_1 \xi_2 - \alpha_2 \xi_1) = a \cdot \Gamma_x$, $x \cdot a = (0, 0, \xi_1 \alpha_2 - \xi_2 \alpha_1) = a \Delta_x$, where

$$(54) \quad \Gamma_x = -\Delta_x = \begin{pmatrix} 0 & 0 & \xi_2 \\ 0 & 0 & -\xi_1 \\ 0 & 0 & 0 \end{pmatrix}.$$

This algebra is associative since

$$(55) \quad \Gamma_x \Delta_y = \begin{pmatrix} 0 & 0 & \xi_2 \\ 0 & 0 & -\xi_1 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 & -\eta_2 \\ 0 & 0 & \eta_1 \\ 0 & 0 & 0 \end{pmatrix} = 0 = \Delta_y \Gamma_x.$$

We let

$$(56) \quad \Lambda = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

and define

$$(57) \quad \Gamma_x^{(0)} = \Lambda \Gamma_x = \begin{pmatrix} 0 & 0 & \xi_1 \\ 0 & 0 & \xi_2 \\ 0 & 0 & 0 \end{pmatrix}, \quad \Delta_x^{(0)} = \Delta_{x\Lambda} = \begin{pmatrix} 0 & 0 & \xi_1 \\ 0 & 0 & \xi_2 \\ 0 & 0 & 0 \end{pmatrix}.$$

Then $\Gamma_x^{(0)} \Delta_x^{(0)} = \Delta_x^{(0)} \Gamma_x^{(0)} = 0$ as before. But we have then defined an isotope \mathfrak{A}_0 of \mathfrak{A} with $(e_1, e_1) = e_1 \Gamma_x e_1^{(0)} = e_3$. It is not equivalent to \mathfrak{A} since in \mathfrak{A} the square of every a is $a \Gamma_a = (0, 0, \alpha_1 \alpha_2 - \alpha_2 \alpha_1) = 0$.

15. Isotopes with a prescribed unity quantity

Let \mathfrak{N} be a linear subspace of order n over \mathfrak{F} of $(\mathfrak{F})_n$ and I be in \mathfrak{N} . Then we have seen that if f is any non-zero vector of \mathfrak{N} and the linear mapping

$$(58) \quad R \rightarrow fR$$

of \mathfrak{N} on \mathfrak{N} is one-to-one the algebra \mathfrak{A} with $R(\mathfrak{A}) = \mathfrak{N}$ and which is defined by

$$(59) \quad fS \cdot fR = f \cdot SR \quad (S \text{ in } (\mathfrak{F})_n, R \text{ in } \mathfrak{N})$$

has f as left unity quantity, $L_f = I$. But also if $fR = x$ we have $R = R_x$. Since \mathfrak{N} contains I we have $fI = f$, $I = R_f$, f is the unity quantity of \mathfrak{A} .

Conversely if \mathfrak{A} has f as its unity quantity we have $f \cdot x = fR_x = x$ and the linear mapping $R_x \rightarrow fR_x$ is one-to-one and is such that $I = R_f$ is in $\mathfrak{N} = R(\mathfrak{A})$.

Let us now assume that \mathfrak{A} is a prescribed algebra with a unity quantity e so that $R(\mathfrak{A})$ contains I . We now let P and C be non-singular linear transformations and G in $R(\mathfrak{A})$ have the property that

$$(60) \quad PGC = I.$$

Then $\mathfrak{N} = PR(\mathfrak{A})C$ contains I and if we let f be any vector such that the linear mapping

$$N \rightarrow fN$$

of \mathfrak{N} on \mathfrak{F} is one-to-one we define an algebra \mathfrak{A}_0 with $R(\mathfrak{A}_0) = \mathfrak{N}$ by $fS \cdot fN = fSN$ for every S in $(\mathfrak{F})_n$ and N in \mathfrak{N} and have seen that f is its unity quantity. It is then desirable to have

THEOREM 16. *The algebra \mathfrak{A}_0 with unity quantity f defined above is an isotope of \mathfrak{A} with f as unity quantity.*

For $fN = x$ implies that $N = R_x^{(0)} = PR_xC$. Write $g = fP$ and have $x = gR_xC = (g \cdot z)C = zL_gC$. If L_g is singular so is L_gC , that is the mapping of N on $fN = x$ is singular. It follows that $N \rightarrow fN$ is non-singular if and only if $g = fP$ is a left non-singular quantity of \mathfrak{A} . We put $Q = C^{-1}(L_g)^{-1}$ and have $z = xQ$, $R_x^{(0)} = PR_{xQ}C$ as desired.

16. Commutative isotopes

An algebra \mathfrak{A} is commutative if and only if $a \cdot x = aR_x = x \cdot a = aL_x$, that is,

$$(61) \quad R_x = L_x$$

for every x of \mathfrak{A} . Let \mathfrak{A}_0 be a principal isotope of an algebra \mathfrak{A} (which may or may not be commutative) so that \mathfrak{A}_0 is commutative if and only if $PR_{xQ} = QL_{xP}$ for every x . Put $y = xP$ and have $xQ = yP^{-1}Q = yS$ where $S = P^{-1}Q$. Thus \mathfrak{A}_0 is commutative if and only if

$$(62) \quad R_{yS} = SL_y$$

for every y of \mathfrak{A} .

Suppose now that \mathfrak{A} has a left unity quantity e , that is, $e \cdot x = x$ for every x of \mathfrak{A} , $L_e = I$. Then (62) implies that $S = R_f$. Thus f is a right non-singular quantity of \mathfrak{A} . Moreover $yS = yR_f = y \cdot f$ and (62) is equivalent to $x \cdot (yS) = xR_fL_y = y \cdot (x \cdot f)$, that is, to

$$(63) \quad x \cdot (y \cdot f) = y \cdot (x \cdot f)$$

for every x and y of \mathfrak{A} .

Conversely let P be any non-singular linear transformation, f be any right non-singular quantity of \mathfrak{A} and put $Q = PR_f$ so that $S = R_f = P^{-1}Q$ and (63) implies (62). We have proved

THEOREM 17. *Let \mathfrak{A} be an algebra with a left unity quantity, f range over all right non-singular quantities of \mathfrak{A} such that (63) holds for every x and y of \mathfrak{A} . Then the principal isotopes of \mathfrak{A} defined by $R_x^{(0)} = PR_{xPR_f}$, for P any non-singular quantity of $(\mathfrak{F})_n$, are commutative algebras to one of which every commutative isotope of \mathfrak{A} is equivalent.*

17. Division algebras

An algebra \mathfrak{A} is called a *division algebra* if it has no (right and hence no left) divisors of zero. Then every non-zero quantity of \mathfrak{A} is both left and right non-singular and we apply Theorem 7 with $g = h \neq 0$ to obtain as an immediate consequence.¹⁰

¹⁰ This result was also obtained by Steenrod. His proof was necessarily more complicated as he did not have Theorems 5 and 7.

THEOREM 18. *Every division algebra is isotopic to a division algebra with a unity quantity.*

Non-associative division algebras do not have many of the properties of associative division algebras. In particular the right minimum function of a quantity of a division algebra may be reducible. Let us note now that a division algebra \mathfrak{A} has no right ideals other than \mathfrak{A} or zero. For otherwise we would have $b \cdot x$ in a right ideal \mathfrak{B} for every b of \mathfrak{B} and x of \mathfrak{A} whereas $b \cdot x = a$ has the solution $x = a(L_b)^{-1}$ for every a of \mathfrak{A} . Similarly \mathfrak{A} has no left ideals other than \mathfrak{A} or zero.

If \mathfrak{B} is a division subalgebra of an algebra \mathfrak{A} and the unity quantity of \mathfrak{A} is in \mathfrak{B} we may prove that if u is any quantity in \mathfrak{A} and not in \mathfrak{B} the linear spaces \mathfrak{B} , $u\mathfrak{B}$ are supplementary in their sum. For otherwise $u \cdot b = b_0$ for non-zero b and b_0 in \mathfrak{B} , $u = b_0(R_b)^{-1}$. But the equation $xb = b_0$ has the solution $x = b_0(R_b)^{-1}$ in the division algebra \mathfrak{B} and u is in \mathfrak{B} , a contradiction.

The process above is used for associative algebras to prove the theorem that the order of \mathfrak{B} divides the order of \mathfrak{A} under the hypothesis just stated. The usual proof of this result requires that if $u_1\mathfrak{B} + \dots + u_r\mathfrak{B} = \mathfrak{B}$, is a supplementary sum of linear spaces not containing u then so is the sum of \mathfrak{B} , and $u\mathfrak{B}$. Thus in particular we need to show that if u is not in $v\mathfrak{B}$ no non-zero quantity of $u\mathfrak{B}$ is in $v\mathfrak{B}$. But $u \cdot b = v \cdot b_0$ then $u = (v \cdot b_0)R_b^{-1} = vR_{b_0}(R_b)^{-1}$. However we cannot conclude that this latter quantity is in $v\mathfrak{B}$ and thus our proof breaks down. We leave the question as to the validity of this theorem as an unsolved problem.

If \mathfrak{A} is a division algebra every R_x defined for $x \neq 0$ is non-singular and $fR_x = 0$ if and only if $f = 0$. Thus the mapping $R \rightarrow fR$ of $R(\mathfrak{A})$ on \mathfrak{A} is non-singular for every $f \neq 0$. Moreover so is the mapping $PRC \mapsto fPRC$ for every non-singular P and C . By Theorem 16 we have

THEOREM 19. *Let f be any non-zero quantity of a division algebra \mathfrak{A} and P and Q be any non-singular linear transformations such that $PRQ = I$ for some R of $R(\mathfrak{A})$. Then the algebra \mathfrak{A}_0 defined by $(a, fS) = aS$ for every S of $PR(\mathfrak{A})Q$ is an isotope of \mathfrak{A} with f as unity quantity.*

If $\phi(\lambda)$ is the right minimum function of a quantity b in a division algebra \mathfrak{A} with a unity quantity e and $\phi(\lambda)$ is reducible and of degree $t > 1$ it cannot have a linear factor. For otherwise $\phi(\lambda) = \psi(\lambda)[\lambda - \alpha]$ and $\phi_R(b) = [e\psi(R_b)](R_b - \alpha I) = \psi_R(b) \cdot (b - \alpha e) = 0$ which is impossible since $\psi_R(b) \neq 0$, $b - \alpha e \neq 0$. However it is possible that $\phi(\lambda) = \psi(\lambda)(\lambda^2 + \alpha\lambda + \beta)$ since then $\phi_R(b) = [e\psi(R_b)](R_b^2 - \alpha R_b + \beta I) \neq [\psi(b)] \cdot (b^2 - \alpha b + \beta e)$ since in general $R_b^2 \neq R_{b^2}$. It then becomes of interest to ask whether or not any quantity of a division algebra has irreducible right minimum function. It is not easy to answer this but we may prove instead

THEOREM 20. *Let \mathfrak{A} be a division algebra with a unity quantity e over \mathfrak{F} and let b in \mathfrak{A} be not in \mathfrak{F} . Then there exists an isotope \mathfrak{A}_0 of \mathfrak{A} such that \mathfrak{A}_0 has a unity quantity, $R(\mathfrak{A}_0) = R(\mathfrak{A})$, the right minimum function of b in \mathfrak{A}_0 is irreducible. For it suffices to assume that the right minimum function $\phi(\lambda)$ of b is reducible,*

$\phi(\lambda) = \pi(\lambda)\psi(\lambda)$ where $\psi(\lambda)$ is irreducible and has degree $t > 1$. Then $\phi_R(b) = e\phi(R_b) = e\pi(R_b) \cdot \psi(R_b) = f\psi(R_b) = 0$ where $f = e\pi(R_b) = \pi_R(b) \neq 0$. We pass to the isotope \mathfrak{A}_0 defined as in Theorem 19 for $P = Q = I$, $R(\mathfrak{A}_0) = R(\mathfrak{A})$ and have f as the unity quantity of \mathfrak{A}_0 . Then $f\psi(R_b) = \psi_R(b) = 0$ in \mathfrak{A}_0 . By Theorem 2 the right minimum function of b divides $\psi(\lambda)$ and must coincide with this irreducible polynomial.

The result just obtained implies that every division algebra of order $n > 1$ over the field \mathfrak{R}' of all real numbers is isotopic to a division algebra with a unity quantity e and containing a quantity b such that $b^2 = -e$. Moreover it is clear that every division algebra \mathfrak{A} of order $n > 2$ over \mathfrak{R}' is central simple. For otherwise we could write \mathfrak{A} as a division algebra over its center $\mathfrak{C} \neq \mathfrak{R}'$, \mathfrak{C} must be $\mathfrak{R}'(i)$ for $i^2 = -1$, \mathfrak{A} over \mathfrak{C} has an isotope \mathfrak{A}_0 over \mathfrak{C} such that b in \mathfrak{A}_0 has $\lambda^2 + 1$ as (right) minimum function. But $\lambda^2 + 1 = (\lambda + i)(\lambda - i)$ in \mathfrak{C} contrary to the proof above.

18. Subalgebras of isotopes

The problem of finding in a division algebra a quantity whose right minimum function is irreducible is an instance of the problem of determining whether an algebra \mathfrak{A} has a certain type of subalgebra. In particular we may ask whether or not a given algebra \mathfrak{A} has any proper subalgebras. A criterion that this be the case was given in Lemma 2 and we wish now to propose the question as to whether a principal isotope of \mathfrak{A} has subalgebras of the same order as those of \mathfrak{A} . By Lemma 4 we have $\mathfrak{B} = \mathfrak{A}E$ is a subalgebra of \mathfrak{A} whose order is the rank of $E \neq 0$ if and only if $ER_y = ER_yE$, $EL_y = EL_yE$ for $y = xE$ and every x of \mathfrak{A} . Now $R_x^{(0)} = PR_{xQ}$, $L_x^{(0)} = QL_{xP}$ and $\mathfrak{A}E_0$ is a subalgebra of \mathfrak{A}_0 if and only if $E_0R_z^{(0)} = E_0R_z^{(0)}E_0$, $E_0L_z^{(0)} = E_0L_z^{(0)}E_0$ for every x of \mathfrak{A} where $z = xE_0$.

The problem just proposed does not appear to have a simple solution for arbitrary algebras. However we should observe that if \mathfrak{A}_0 has a unity quantity then $P = (R_h)^{-1}$, $Q = (L_h)^{-1}$ and if g and h are in \mathfrak{B} the linear space \mathfrak{B} is a subalgebra of \mathfrak{A}_0 as well as of \mathfrak{A} . For P and Q are in $T(\mathfrak{B}, \mathfrak{A}) = ET(\mathfrak{B}, \mathfrak{A})E$ and $ER_y^{(0)} = EPR_{yQ} = EPER_{yQ} = EPER_{yQ}E = ER_y^{(0)}$ since $yQ = xEQ = xEQE = yQE$ is in \mathfrak{B} . Similarly $EL_y^{(0)} = EL_y^{(0)}E$. We shall not study the general question further except to note that if E_0 has the same rank as E it has the form $H^{-1}EH$ and it may be seen that $\mathfrak{B}_0 = \mathfrak{A}E_0$ is a subalgebra of \mathfrak{A}_0 if and only if \mathfrak{B} is a subalgebra of the isotope \mathfrak{A}_1 defined by $R_x^{(1)} = HR_{xH}^{-1}$ and equivalent to \mathfrak{A}_0 .

19. Special properties

An algebra \mathfrak{A} is said to be alternative if $(a \cdot x) \cdot x = a \cdot (x \cdot x)$, $x \cdot (x \cdot a) = (x \cdot x) \cdot a$ for every x and a of \mathfrak{A} . Then \mathfrak{A} is alternative if and only if

$$(64) \quad R_{x^2} = (R_x)^2, \quad L_{x^2} = (L_x)^2$$

for every x of \mathfrak{A} .

It follows from (64) that $x \cdot x^2 = x(R_x)^2 = (xR_x)R_x = x^2 \cdot x$. Suppose then

that $R_{x^k} = (R_x)^k$ for all right powers $k = 1, 2, \dots, t$ and that $x \cdot x^k = x^k \cdot x$ for $k = 1, \dots, t$. Then we put $y = x + x^t$ and have $y^2 = x^2 + (x^t)^2 + x \cdot x^t + x^t \cdot x = x^2 + (x^t)^2 + 2x^{t+1}$. But $R_y = R_x + R_{x^t} = R_x + (R_x)^t$, $(R_y)^2 = (R_x)^2 + 2(R_x)^{t+1} + (R_x)^{2t} = R_{x^2} + R_{x^{t+1}} + 2(R_x)^{t+1}$. It follows that $2R_{x^{t+1}} = 2(R_x)^{t+1}$ and that $(R_x)^{t+1} = R_{x^{t+1}}$ if the characteristic of \mathfrak{F} is not two. But then $x \cdot x^{t+1} = x(R_x)^{t+1} = (xR_{x^t})R_x = x^{t+1} \cdot x$. This completes our induction and proves that $R_{x^k} = (R_x)^k$ for every k .

We see that consequently $x^k \cdot x^t = xR_{x^{t+k-1}} = x^{t+k}$ so that all powers of x are right powers, $(x^s \cdot x^t) \cdot x^k = x^{s+t+k} = x^s \cdot (x^t \cdot x^k)$. It follows that the algebra $\mathfrak{F}[x]$ of all right polynomials $\phi_R(x)$ is the associative algebra of all polynomials $\phi(x) = \phi_R(x) = \phi_L(x)$ and $R_{\phi(x)} = \phi(R_x)$, $L_{\phi(x)} = \phi(L_x)$.

We now propose the problem of determining the principal isotopes of an algebra \mathfrak{A} which are alternative. This occurs if and only if $(R_x^{(0)})^2 = R_x^{(0)}$, $(L_x^{(0)})^2 = L_x^{(0)}$ where $z = xR_x^{(0)} = xL_x^{(0)}$. But then we must have

$$R_{xQ}PR_{xQ} = R_{xQ}, \quad L_{xP}QL_{xP} = L_{xP}.$$

Replace xQ by x and thus $zQ = xQL_{xP}Q$ by $(xQ^{-1}P \cdot x)Q$ and similarly replace xP by x and thus $zP = xPR_{xQ}P$ by $(xP^{-1}Q \cdot x)P$. Then we see that \mathfrak{A}_0 is alternative if and only if

$$R_xPR_x = R_u, \quad L_xQL_x = L_v,$$

for every x of \mathfrak{A} , where

$$u = (xQ^{-1}P \cdot x)Q, \quad v = (xP^{-1}Q \cdot x)P,$$

and the indicated products are those in \mathfrak{A} . It is of particular interest, of course, to study the case where we assume also that \mathfrak{A} is alternative.

An algebra \mathfrak{A} is called a *Lie algebra* if $a \cdot x = -x \cdot a$, $a \cdot (x \cdot y) + y \cdot (a \cdot x) + x \cdot (y \cdot a) = 0$ for every a, x, y of \mathfrak{A} . Then $L_x = -R_x$, $aR_{x \cdot y} + aR_xL_y + aL_yL_x = 0$, $a[L_{x \cdot y} - (R_xR_y - R_yR_x)] = 0$, \mathfrak{A} is a Lie algebra if and only if

$$(65) \quad L_x = -R_x, \quad R_{x \cdot y} = R_xR_y - R_yR_x \quad (x, y \text{ in } \mathfrak{A}).$$

We propose again the question as to whether a principal isotope of an algebra \mathfrak{A} is a Lie algebra and see that this occurs if and only if

$$PR_{xQ} = -QL_{xP}, \quad PR_{xQ}PR_{yQ} - PR_{yQ}PR_{xQ} = PR_z$$

where $z = (x, y)Q = yQL_{xP}Q$. Replace xQ by x , yQ by y and thus xP by xC for $C = Q^{-1}P$, z by $yL_{xC}Q = (xC \cdot y)Q$. Then \mathfrak{A}_0 is a Lie algebra if and only if

$$(66) \quad L_{xC} = -CR_x, \quad R_xPR_y - R_yPR_x = R_{(xC \cdot y)Q}.$$

The problem of determining the principal Lie isotopes of simple Lie algebras is being studied.¹¹

¹¹ This problem is the topic of study of a doctoral dissertation at the University of Chicago. We also wish to mention here that Mr. W. Carter in his Master's dissertation has classified all real division algebras of order four and degree two into classes of algebras

Let us conclude these remarks with some observations which will be important for the study of simple algebras. We define the *center*¹² of any algebra \mathfrak{A} over \mathfrak{F} to be the set \mathfrak{Z} of all quantities z of \mathfrak{A} such that $R_z = L_z$ is commutative with every R_x of $R(\mathfrak{A})$ and L_x of $L(\mathfrak{A})$. Then z is in \mathfrak{Z} if and only if

$$z \cdot a = a \cdot z, \quad z \cdot (a \cdot x) = (z \cdot a) \cdot x, \quad a \cdot (z \cdot x) = (a \cdot z) \cdot x, \quad (a \cdot x) \cdot z = a \cdot (x \cdot z)$$

for every a and x of \mathfrak{A} . It is easily shown that \mathfrak{Z} is zero or an associative subalgebra of \mathfrak{A} . Moreover if \mathfrak{A} has a unity quantity e the set $\mathfrak{Y} = e\mathfrak{F}$ of all $e\alpha$ for α in \mathfrak{F} is a subalgebra of order one over \mathfrak{F} of \mathfrak{Z} , $e\mathfrak{F}$ is equivalent to \mathfrak{F} and is a field.¹³

Let us define a new operation of scalar product on $\mathfrak{A}\mathfrak{Y}$ to \mathfrak{A} by writing $(a, y) = a \cdot y = a \cdot e\alpha = a\alpha$ for every a of \mathfrak{A} and α of \mathfrak{F} . Then \mathfrak{A} is an algebra over \mathfrak{Y} with respect to this operation. It is clear that this is a change in our representation of \mathfrak{A} as a linear space over a field and is not a change in \mathfrak{A} .

If \mathfrak{A} is a simple algebra of order n over \mathfrak{F} we have seen that \mathfrak{A} is a central simple algebra of order s over its transformation center \mathfrak{C} , \mathfrak{C} is a field of degree t over \mathfrak{F} , $n = st$. Let \mathfrak{A} have e as its unity quantity so that, as above, $e\mathfrak{C}$ is a subalgebra over \mathfrak{C} of \mathfrak{A} and is equivalent to \mathfrak{C} . But then $e\mathfrak{C}$ is a field of degree t over $e\mathfrak{F}$, $e\mathfrak{C}$ is a subalgebra of \mathfrak{A} of order t over \mathfrak{F} . Moreover it is easy to verify that $e\mathfrak{C}$ is contained in the center \mathfrak{Z} of \mathfrak{A} . But the quantities of \mathfrak{Z} are quantities z such that R_z is in \mathfrak{C} , $e \cdot z = eR_z = z$ is in $e\mathfrak{C}$. This proves that $e\mathfrak{C}$ is the center of every simple algebra \mathfrak{A} with e as unity quantity and \mathfrak{C} as its transformation center. If we express \mathfrak{A} as an algebra over \mathfrak{C} it is central simple and consequently we may express \mathfrak{A} as a central simple algebra over the associative subalgebra of \mathfrak{A} which is its center \mathfrak{Z} .

It is desirable to note that if \mathfrak{A} and \mathfrak{A}_0 are principal isotopes we have $T(\mathfrak{A}) = T(\mathfrak{A}_0)$ and hence $\mathfrak{C} = \mathfrak{C}(\mathfrak{A}) = \mathfrak{C}(\mathfrak{A}_0)$, \mathfrak{A} and \mathfrak{A}_0 have the same transformation center. If e and e_0 are corresponding unity quantities we have $e\mathfrak{C}$ equivalent over \mathfrak{F} to $e_0\mathfrak{C}$ and thus we have shown that *isotopic simple algebras with unity quantities have equivalent centers*. It is important to observe that, while the center of an associative simple algebra \mathfrak{A} is its \mathfrak{A} -centralizer, this may not be the case when \mathfrak{A} is not associative.

UNIVERSITY OF CHICAGO

with respect to isotopy. He has also shown that *every real division algebra of order four is isotopic to an algebra of degree four*, that is, containing a quantity whose right minimum function has degree four and is thus reducible. Moreover *there exist real division algebras of degree and order four not isotopic to algebras of degree two*.

¹² We have now used the terms *center* instead of *centrum* and *central* in place of *normal*. A change of terminology of this kind has long seemed very desirable to many algebraists.

¹³ In particular \mathfrak{A} may be commutative and may yet be a central simple algebra. For example the set of all r -rowed real symmetric square matrices forms a commutative central simple algebra with respect to the product operation $a \cdot x = \frac{1}{2}(ax + xa)$, ax and xa the ordinary matrix products.

NON-ASSOCIATIVE ALGEBRAS

II. New Simple Algebras¹

By A. A. ALBERT

(Received January 30, 1942)

1. Introduction

In the second part of our study of non-associative algebras we shall give an iterative construction of new simple algebras with a unity quantity.* All previous constructions² of this type have used groups of automorphisms or anti-automorphisms and the great generality of our definition will lie precisely in that we shall be able to use instead almost³ arbitrary multiplicative groups of non-singular linear transformation.

We shall begin our exposition with a preliminary discussion of (non-associative) separable algebras, that is algebras \mathfrak{A} with a unity quantity e such that every scalar extension of \mathfrak{A} is a direct sum of simple algebras. Let \mathfrak{G} be any finite multiplicative group of non-singular linear transformations S on \mathfrak{A} such that $eS = e$ and \mathfrak{H} be any subset of \mathfrak{G} containing the identity transformation. We define an *extension set* \mathfrak{g} to be a set of non-singular quantities $g_{S,T}$ in \mathfrak{A} for every S and T in \mathfrak{G} . Then we shall construct a corresponding *crossed extension* $\mathfrak{E} = (\mathfrak{A}, \mathfrak{G}, \mathfrak{H}, \mathfrak{g})$ which is a certain algebra having e as its unity quantity.

For every separable \mathfrak{A} we shall give conditions that the crossed extensions shall be simple (or central simple) algebras. If \mathfrak{H} is the identity group \mathfrak{E} is simple whenever \mathfrak{A} is, \mathfrak{E} is central simple whenever \mathfrak{A} is. These latter algebras include the so-called *Cayley algebras*. Our algebras are associative only when $\mathfrak{G} = \mathfrak{H}$ is a group of automorphisms and our definition then includes that of crossed products.⁴

The crossed products are associative central simple algebras of order r^2 , and for each such algebra we may use an explicit process to give a set of corresponding central simple algebras of order r^t for any integer $t > 1$. These algebras are not associative for $t > 2$. In particular we then have generalized cyclic algebras. Another explicit construction will connect every central simple

¹ Presented to the Society February 28, 1942.

² Automorphisms were necessarily used in the constructions of associative algebras of L. E. Dickson for which see my *Structure of Algebras*, Chapter V, Chapter XI, pp. 182-8, and bibliographical references [141], [145]. Cayley algebras were generalized in my "Quadratic forms permitting composition" these *Annals*, vol. 41, pp. 161-77, to algebras of order 2^t obtained by the process given here where the group consists of the identity automorphism and an antiautomorphism of order two.

³ The special restrictions will reduce to the property that the transformations leave the unity quantity unaltered in the most important cases.

⁴ For these algebras and the Cayley algebras see the references in footnote 2.

algebra of order n with a crossed extension of it by a group \mathfrak{G} equivalent to any permutation group on m letters.

We shall close our discussion with a list of fundamental unsolved problems in the theory of these new algebras.

2. Decomposition of algebras with a unity quantity

An algebra \mathfrak{A} is said to be *decomposable*⁵ if it is expressible as the supplementary sum $\mathfrak{A}_1 + \cdots + \mathfrak{A}_r$, of at least two subalgebras \mathfrak{A}_i of \mathfrak{A} , such that $a_i a_j = 0$ for a_i in \mathfrak{A}_i , a_j in \mathfrak{A}_j and all $i \neq j$. Then we say that \mathfrak{A} decomposes into the *direct sum* of its *components* \mathfrak{A}_i (which are ideals of \mathfrak{A}), we write

$$(1) \quad \mathfrak{A} = \mathfrak{A}_1 \oplus \cdots \oplus \mathfrak{A}_r,$$

and we call (1) a *decomposition* of \mathfrak{A} . If \mathfrak{A} has no such decomposition we say that \mathfrak{A} is *indecomposable*. It is clear that a decomposition with r components becomes one with $r + 1$ components if we replace a decomposable component $\mathfrak{A}_i = \mathfrak{B} \oplus \mathfrak{C}$ by $\mathfrak{B} \oplus \mathfrak{C}$ in (1). The lowering of orders in such decomposition implies that every \mathfrak{A} has a decomposition (1) with the components indecomposable algebras. It is then natural to ask (as in the theory of associative algebras) whether or not such a decomposition is unique apart from the ordering of the components. We shall give a simple solution below for algebras with a unity quantity.

The *center* of an algebra \mathfrak{A} has been defined⁶ to be the set \mathfrak{Z} of all quantities z of \mathfrak{A} such that the commutative and associative laws, for products in \mathfrak{A} , hold whenever z is one of the factors. Then \mathfrak{Z} is zero or an associative and commutative subalgebra of \mathfrak{A} . When \mathfrak{A} has a unity quantity e the subalgebra $e\mathfrak{A}$ of \mathfrak{A} is in its center \mathfrak{Z} and we call it a *central algebra* if $\mathfrak{Z} = e\mathfrak{A}$. We may now prove

LEMMA 1. *Let \mathfrak{A} be an algebra with a unity quantity e and \mathfrak{Z} be the center of \mathfrak{A} . Then \mathfrak{A} has the form (1) if and only if $e = e_1 + \cdots + e_r$ for pairwise orthogonal idempotents e_i in \mathfrak{Z} such that $\mathfrak{A}_i = \mathfrak{A}e_i$.*

For if \mathfrak{A} has the form (1) we may write every quantity of \mathfrak{A} in the form $a = a_1 + \cdots + a_r$, for the a_i uniquely determined quantities of \mathfrak{A}_i . Then if $x = x_1 + \cdots + x_r$ we have $a \cdot x = a_1 \cdot x_1 + \cdots + a_r \cdot x_r$. Take $x = e = e_1 + \cdots + e_r$ and have $e_i \cdot e_j = 0$ for $i \neq j$, $a \cdot e = a_1 \cdot e_1 + \cdots + a_r \cdot e_r = a$ if and only if $a_i \cdot e_i = a_i$. Similarly $e_i \cdot a_i = a_i$, \mathfrak{A}_i has e_i as its unity quantity, $\mathfrak{A}_i = \mathfrak{A}e_i$. Now $(a \cdot x) \cdot e_i = (a_i \cdot x_i) \cdot e_i = a_i \cdot x_i = (a \cdot e_i) \cdot x_i = a \cdot (x e_i)$ and similar other verifications imply that e_i is in \mathfrak{Z} . Conversely if $e = e_1 + \cdots + e_r$, for pairwise orthogonal idempotents e_i in \mathfrak{Z} , we have $\mathfrak{A} = \mathfrak{A}_1 + \cdots + \mathfrak{A}_r$ for $\mathfrak{A}_i = \mathfrak{A}e_i$ and we have (1).

As a consequence of this result we have

⁵ This term seems much preferable to the term reducible which causes so much confusion if representation theory and linear algebra theory be considered together.

⁶ This was defined at the end of part I of this paper. We shall use the concepts given in that part without any reference.

LEMMA 2. *The algebra \mathfrak{A} of Lemma 1 has a decomposition (1) if and only if its center \mathfrak{Z} has a corresponding decomposition*

$$(2) \quad \mathfrak{Z} = \mathfrak{Z}_1 + \cdots + \mathfrak{Z}_r.$$

Then \mathfrak{Z}_i is the intersection of \mathfrak{A}_i and \mathfrak{Z} and is the center of \mathfrak{A}_i .

For e is in \mathfrak{Z} and Lemma 1 implies that from (1) we have (2) with

$$(3) \quad \mathfrak{Z}_i = \mathfrak{Z}e_i$$

and conversely (2) and (3) imply (1) with $\mathfrak{A} = \mathfrak{A}e_i$. Then \mathfrak{Z}_i is in both \mathfrak{Z} and \mathfrak{A}_i , and is the intersection of \mathfrak{Z} and \mathfrak{A}_i by (2). If z_i is in the center of \mathfrak{A}_i then $z_i \cdot a_j = a_j \cdot z_i = 0$ for a_j in \mathfrak{A}_j and $j \neq i$, z_i is in \mathfrak{Z}_i , \mathfrak{Z}_i is the center of \mathfrak{A}_i .

Lemma 2 clearly implies

LEMMA 3. *An algebra \mathfrak{A} with a unity quantity is indecomposable if and only if its center is indecomposable.*

We then have

LEMMA 4. *The decomposition of an algebra with a unity quantity as a direct sum (1) of indecomposable components is unique apart from the arrangement of the components.*

This follows from Lemmas 2, 3, and the associative case of the result we are proving. This latter result is proved by the use of the following lemma which may then be used to prove Lemma 4.

LEMMA 5. *Let an algebra \mathfrak{A} with a unity quantity have a decomposition (1) so that $\mathfrak{A}_i = \mathfrak{A}e_i$ with e_i an idempotent of \mathfrak{A} . Then every right, left or two-sided ideal \mathfrak{B} of \mathfrak{A} is the direct sum*

$$(4) \quad \mathfrak{B} = \mathfrak{B}_1 \oplus \cdots \oplus \mathfrak{B}_r,$$

where $\mathfrak{B}_i = \mathfrak{B}e_i$ is the intersection of \mathfrak{B} and \mathfrak{A}_i , \mathfrak{B}_i is correspondingly a right, left, or two-sided ideal of \mathfrak{A}_i .

The proof of this result involves the use of the associative law only for products $a \cdot b \cdot c$ with a factor in the center, and thus the proof which has been given in the associative case is valid without change. We shall not repeat it here.

3. Absolute indecomposability

If \mathfrak{Z} is the center of an algebra \mathfrak{A} over \mathfrak{F} and \mathfrak{R} is any scalar extension of \mathfrak{F} the center of $\mathfrak{A}_{\mathfrak{R}}$ is $\mathfrak{Z}_{\mathfrak{R}}$. For it is clear that $\mathfrak{Z}_{\mathfrak{R}}$ is contained in the center \mathfrak{Z}_0 of $\mathfrak{A}_{\mathfrak{R}}$. Let then z_0 be in \mathfrak{Z}_0 so that we may write $z_0 = z_1\xi_1 + \cdots + z_r\xi_r$ where the z_i are in \mathfrak{A} and ξ_i in \mathfrak{R} are such that a sum $a_1\xi_1 + \cdots + a_r\xi_r = 0$ for the a_i in \mathfrak{A} only when the a_i are all zero. Then if a is in \mathfrak{A} we have $a \cdot z_0 - z_0 \cdot a = (a \cdot z_1 - z_1 \cdot a)\xi_1 + \cdots + (a \cdot z_r - z_r \cdot a)\xi_r = 0$, and $a \cdot z_i - z_i \cdot a = 0$. If also x is in \mathfrak{A} we compute $a \cdot (x \cdot z_0) - (a \cdot x) \cdot z_0$ and other similar products, and see that the z_i are in \mathfrak{Z} , z_0 is in $\mathfrak{Z}_{\mathfrak{R}}$, $\mathfrak{Z}_0 = \mathfrak{Z}_{\mathfrak{R}}$.

An algebra \mathfrak{A} over \mathfrak{F} may be indecomposable but there may exist a scalar extension \mathfrak{R} of \mathfrak{F} such that $\mathfrak{A}_{\mathfrak{R}}$ is decomposable. Thus we call \mathfrak{A} *absolutely*

*indecomposable*⁷ if $\mathfrak{A}_{\mathfrak{R}}$ is indecomposable for every \mathfrak{R} . Moreover a decomposition (1) of a decomposable algebra will be called an *absolute decomposition* if the components \mathfrak{A}_i are all absolutely indecomposable. Lemma 2 and the result above then imply

LEMMA 6. *An algebra \mathfrak{A} is absolutely indecomposable if and only if its center \mathfrak{Z} is absolutely indecomposable.*

LEMMA 7. *A decomposition (1) is an absolute decomposition if and only if the center of each component \mathfrak{A}_i is absolutely indecomposable.*

We may also use Lemma 4 to obtain

LEMMA 8. *Let \mathfrak{A} be an algebra with a unity quantity, \mathfrak{R} and \mathfrak{R}_0 be scalar extensions of \mathfrak{F} such that*

$$(5) \quad \mathfrak{A}_{\mathfrak{R}} = \mathfrak{A}_1 + \cdots + \mathfrak{A}_r, \quad \mathfrak{A}_{\mathfrak{R}_0} = \mathfrak{B}_1 + \cdots + \mathfrak{B}_s$$

for absolutely indecomposable \mathfrak{A}_i and \mathfrak{B}_j . Then $r = s$ and, if we imbed \mathfrak{R} and \mathfrak{R}_0 in a scalar extension \mathfrak{R}_1 of \mathfrak{F} , there is a permutation j_1, \dots, j_r of $1, 2, \dots, r$ such that $(\mathfrak{A}_i)_{\mathfrak{R}_1} = (\mathfrak{B}_{j_i})_{\mathfrak{R}_1}$.

For $\mathfrak{A}_{\mathfrak{R}_1} = (\mathfrak{A}_{\mathfrak{R}})_{\mathfrak{R}_1} = (\mathfrak{A}_1)_{\mathfrak{R}_1} \oplus \cdots \oplus (\mathfrak{A}_r)_{\mathfrak{R}_1} = (\mathfrak{A}_{\mathfrak{R}_0})_{\mathfrak{R}_1} = (\mathfrak{B}_1)_{\mathfrak{R}_1} \oplus \cdots \oplus (\mathfrak{B}_s)_{\mathfrak{R}_1}$. Our result then follows from Lemma 4.

The center of a simple algebra \mathfrak{A} with a unity quantity is a field and if separable is indecomposable only if its degree is one, \mathfrak{A} is central. Thus we have

LEMMA 9. *A simple algebra with a unity quantity over \mathfrak{F} and separable center is absolutely indecomposable if and only if it is central simple over \mathfrak{F} .*

4. Semi-simple algebras

We shall call an algebra \mathfrak{A} over \mathfrak{F} a *semi-simple algebra* if it has a unity quantity e and is the direct sum (1) of simple components \mathfrak{A}_i . Then we have seen in Lemmas 1, 2 that \mathfrak{A}_i has a unity quantity e_i such that $e = e_1 + \cdots + e_r$, the center \mathfrak{Z}_i of \mathfrak{A}_i is a field, the center \mathfrak{Z} of \mathfrak{A} is the direct sum of the \mathfrak{Z}_i . We shall call \mathfrak{A} *separable* if $\mathfrak{A}_{\mathfrak{R}}$ is semi-simple for every scalar extension \mathfrak{R} .

Let the center \mathfrak{Z} of a simple algebra \mathfrak{A} over \mathfrak{F} and with a unity quantity e be a separable field. Then if \mathfrak{R} is any scalar extension of \mathfrak{F} the algebra $\mathfrak{Z}_{\mathfrak{R}} = \mathfrak{Z}_1 \oplus \cdots \oplus \mathfrak{Z}_r$, where \mathfrak{Z}_i is a separable field over \mathfrak{R} equivalent over \mathfrak{F} to a composite of \mathfrak{Z} and \mathfrak{R} . If e_i is its unity quantity the algebra \mathfrak{Z}_i contains $\mathfrak{Z}e_i$ which is a field over \mathfrak{F} equivalent over \mathfrak{F} to \mathfrak{Z} under the mapping $z \cdot e_i \rightarrow z$. We let u_1, \dots, u_s be a basis of \mathfrak{A} over \mathfrak{Z} and $u_{\sigma} \cdot u_j = \sum_{k=1}^s u_k z_{\sigma j k}$ for the $z_{\sigma j k}$ in \mathfrak{Z} and see that $u_1 \cdot e_i, \dots, u_s \cdot e_i$ are a basis of $\mathfrak{A}e_i$ over $\mathfrak{Z}e_i$, $(u_{\sigma} \cdot e_i) \cdot (u_j \cdot e_i) = \sum_{k=1}^s (u_k \cdot e_i)(z_{\sigma j k} \cdot e_i)$. Then we have a corresponding decomposition $\mathfrak{A}_{\mathfrak{R}} = \mathfrak{A}_1 + \cdots + \mathfrak{A}_r$ where $\mathfrak{A}_i = (\mathfrak{A}_{\mathfrak{R}})e_i = (\mathfrak{A}e_i)_{\mathfrak{R}}$. But the linear mapping $a \rightarrow a \cdot e_i$ is clearly an equivalence over \mathfrak{F} of \mathfrak{A} and $\mathfrak{A}e_i$, $\mathfrak{Z}e_i$ is the center of $\mathfrak{A}e_i$, \mathfrak{A}_i is a simple algebra with center \mathfrak{Z}_i over \mathfrak{R} .

Conversely let \mathfrak{A} be simple and separable. If \mathfrak{Z} is not separable it is known that there exists a scalar extension \mathfrak{R} of \mathfrak{F} such that $\mathfrak{Z}_{\mathfrak{R}}$ contains a quantity

⁷ This too seems a desirable terminology.

$y \neq 0$, $y^h = 0$ for some positive integer h . But by hypothesis $\mathfrak{A}_\mathfrak{z} = \mathfrak{A}_1 \oplus \cdots \oplus \mathfrak{A}_r$ where the center of $\mathfrak{A}_\mathfrak{z}$ is a direct sum of fields and is a separable associative algebra $\mathfrak{Z}_\mathfrak{z}$. However y is properly nilpotent in \mathfrak{Z} , a contradiction. We have proved

LEMMA 10. *A simple algebra with a unity quantity is separable if and only if its center is a separable field.*

We also clearly have

LEMMA 11. *An algebra \mathfrak{A} with a unity quantity is separable if and only if it is a direct sum of separable simple algebras.*

By a well known property of separable fields we have

LEMMA 12. *Let \mathfrak{A} be separable. Then there exists a scalar extension \mathfrak{R} such that $\mathfrak{A}_\mathfrak{z}$ is a direct sum $\mathfrak{A}_1 \oplus \cdots \oplus \mathfrak{A}_r$ of central simple algebras \mathfrak{A}_i over \mathfrak{R} . This is an absolute decomposition of $\mathfrak{A}_\mathfrak{z}$ and r is the order above \mathfrak{F} of the center of \mathfrak{A} .*

5. Extending groups of linear transformations

If u_1, \dots, u_n is a basis of \mathfrak{A} over \mathfrak{F} , any linear transformation G over \mathfrak{F} on \mathfrak{A} has a matrix Γ such that G is given by $(\alpha_1 u_1 + \cdots + \alpha_n u_n)G = \beta_1 u_1 + \cdots + \beta_n u_n$ where

$$(6) \quad (\beta_1, \dots, \beta_n) = (\alpha_1, \dots, \alpha_n)\Gamma.$$

Then if \mathfrak{R} is any scalar extension of \mathfrak{F} we may indicate by $G_\mathfrak{z}$ the linear transformation on $\mathfrak{A}_\mathfrak{z}$ with the same matrix Γ . It is given by the equations above for the α_i and β_i now in \mathfrak{R} . Conversely if the matrix Γ of a linear transformation G_0 on $\mathfrak{A}_\mathfrak{z}$ with respect to a basis u_1, \dots, u_n of the original algebra \mathfrak{A} has elements in \mathfrak{F} then $G_0 = G_\mathfrak{z}$ where the matrix of G in $(\mathfrak{F})_n$ is also Γ .

As in the theory of *groups with operators* we shall consider algebras \mathfrak{A} over \mathfrak{F} with operator sets \mathfrak{G} of linear transformations G over \mathfrak{F} . If \mathfrak{R} is any scalar extension of \mathfrak{F} we shall designate by $\mathfrak{G}_\mathfrak{z}$ the set of all $G_\mathfrak{z}$ on $\mathfrak{A}_\mathfrak{z}$ for G in \mathfrak{G} .

A linear subspace \mathfrak{B} over \mathfrak{F} of \mathfrak{A} will be called \mathfrak{G} -allowable if bG is in \mathfrak{B} for every b of \mathfrak{B} and G of \mathfrak{G} . Then a \mathfrak{G} -allowable ideal of \mathfrak{A} will be called a \mathfrak{G} -ideal and we shall say that \mathfrak{A} is \mathfrak{G} -simple if it has no \mathfrak{G} -ideals other than itself and the zero ideal. Finally we shall say that \mathfrak{A} is \mathfrak{G} -central if every scalar extension $\mathfrak{A}_\mathfrak{z}$ of \mathfrak{A} is $\mathfrak{G}_\mathfrak{z}$ -simple.

We shall restrict all further attention to subgroups \mathfrak{G} of the multiplicative group of all non-singular linear transformations on \mathfrak{A} and shall call such a group \mathfrak{G} an *extending^{*} group* for \mathfrak{A} if the unity quantity e of \mathfrak{A} has the property that $eG = e$ for every G of \mathfrak{G} . Then every subgroup of an extending group for \mathfrak{A} is an extending group for \mathfrak{A} . Moreover \mathfrak{G} is an extending group for \mathfrak{A} if and only if $\mathfrak{G}_\mathfrak{z}$ is an extending group for $\mathfrak{A}_\mathfrak{z}$ where \mathfrak{R} ranges over all scalar extensions of \mathfrak{F} . We now prove

THEOREM 1. *Let \mathfrak{G} be an extending group for a semi-simple algebra $\mathfrak{A} = \mathfrak{A}_1 \oplus$*

* We shall use this terminology in our theorems so as to diminish the size of the statement of the hypotheses we shall find it necessary to make.

$\cdots \oplus \mathfrak{A}_r$ with simple components \mathfrak{A}_i , and let there exist an $a_i \neq 0$ in \mathfrak{A}_i and a transformation G_i in \mathfrak{G} for each $i = 1, \dots, r$ such that $a_i G_i$ is in \mathfrak{A}_i . Then \mathfrak{A} is \mathfrak{G} -simple, and is \mathfrak{G} -central if the \mathfrak{A}_i are all central simple over \mathfrak{F} .

For Lemma 5 states that every \mathfrak{G} -ideal $\mathfrak{B} = \mathfrak{B}_1 \oplus \cdots \oplus \mathfrak{B}_r$ where the intersection of \mathfrak{B} and \mathfrak{A}_i is the ideal \mathfrak{B}_i of \mathfrak{A}_i . If $\mathfrak{B} \neq 0$ some $\mathfrak{B}_j \neq 0$, $\mathfrak{B}_j = \mathfrak{A}_j$ contains $a_j G_j$. But \mathfrak{G} is a group and \mathfrak{B} is a \mathfrak{G} -ideal only if $(a_j G_j) G_j^{-1} = a_j$ is in \mathfrak{B} . Hence a_j in \mathfrak{A}_i is in \mathfrak{B}_i , $\mathfrak{B}_i = \mathfrak{A}_i$. Then \mathfrak{B} contains every a_i of our hypothesis and also every $a_i G_i$. These are non-zero quantities since $a_i G_i = 0$ implies that $a_i G_i G_i^{-1} = a_i = 0$. They are in \mathfrak{B} and in \mathfrak{A}_i and hence in \mathfrak{B}_i , $\mathfrak{B}_i \neq 0$, $\mathfrak{B}_i = \mathfrak{A}_i$, $\mathfrak{B} = \mathfrak{A}$ is \mathfrak{G} -simple. If every \mathfrak{A}_i is central every $(\mathfrak{A}_i)_\#$ is simple and our proof implies that $\mathfrak{A}_\#$ is $\mathfrak{G}_\#$ -simple, \mathfrak{A} is \mathfrak{G} -central.

THEOREM 2. Let \mathfrak{H} be a subgroup of an extending group \mathfrak{G} for \mathfrak{A} . Then if \mathfrak{A} is \mathfrak{H} -simple it is \mathfrak{G} -simple, and if \mathfrak{A} is \mathfrak{H} -central it is \mathfrak{G} -central.

The next result may be regarded as the trivial case $\mathfrak{H} = [I]$ of Theorem 2.

THEOREM 3. A simple algebra \mathfrak{A} is \mathfrak{G} -simple for every \mathfrak{G} . If \mathfrak{A} is central simple it is \mathfrak{G} -central for every \mathfrak{G} .

Every G of an extending group \mathfrak{G} for an algebra \mathfrak{A} induces a linear mapping $b \rightarrow bG$ of a linear subspace \mathfrak{B} of \mathfrak{A} on $\mathfrak{B}G$. If \mathfrak{H} is any subgroup of \mathfrak{G} such that $\mathfrak{B}H$ is a subset of \mathfrak{B} for every H of \mathfrak{H} then the mappings above are a group of non-singular linear transformations on \mathfrak{B} induced by \mathfrak{H} . We shall use this terminology in the formulation of

THEOREM 4. Let the center of a simple algebra \mathfrak{A} over \mathfrak{F} be a (separable) normal field \mathfrak{Z} and let an extending group \mathfrak{G} for \mathfrak{A} have a subgroup inducing in \mathfrak{Z} its automorphism group \mathfrak{H} . Then \mathfrak{A} is \mathfrak{G} -central.

For if \mathfrak{B} is any $\mathfrak{G}_\#$ -ideal of $(\mathfrak{A})_\#$ the set $\mathfrak{B}_\#$ is a $\mathfrak{G}_\#$ -ideal of $\mathfrak{A}_\#$, where \mathfrak{A} is any scalar extension of \mathfrak{F} containing \mathfrak{A} . But it is well known that \mathfrak{A} may be so chosen that $\mathfrak{Z}_\# = e_1 \mathfrak{A} + \cdots + e_r \mathfrak{A}$ for pairwise orthogonal idempotents e_i , $\mathfrak{A}_\# = \mathfrak{A}_1 \oplus \cdots \oplus \mathfrak{A}_r$ as in Lemma 12, $\mathfrak{A}_i = (\mathfrak{A}_\#) e_i$ central simple. Moreover $e_i = e_i H_i$ for H_i in \mathfrak{H} and hence $e_i = e_i G_i$ for G_i in \mathfrak{G} . By Theorem 1 $\mathfrak{A}_\#$ is \mathfrak{G} -simple, $\mathfrak{B}_\# = 0$ or $\mathfrak{A}_\#$, $\mathfrak{B} = 0$ or \mathfrak{A} , \mathfrak{A} is \mathfrak{G} -central.

COROLLARY I. A normal field \mathfrak{A} over \mathfrak{F} is \mathfrak{G} -central with respect to its (extending) automorphism group \mathfrak{G} .

Theorem 4 may be extended to direct sums and the result stated as

THEOREM 5. Let \mathfrak{A} be a \mathfrak{G} -simple algebra of Theorem 1 such that the center \mathfrak{Z}_i of \mathfrak{A}_i is a normal field over \mathfrak{F} . Then if \mathfrak{G} has subgroups \mathfrak{G}_i for $i = 1, \dots, r$ such that \mathfrak{G}_i induces in \mathfrak{Z}_i its automorphism group, the algebra \mathfrak{A} is \mathfrak{G} -central.

This generalization is a corollary of Theorem 4 and the following

THEOREM 6. Let \mathfrak{A} be a \mathfrak{G} -simple algebra of Theorem 1 and \mathfrak{G} have subgroups \mathfrak{G}_i inducing in \mathfrak{A}_i an extending group \mathfrak{H} such that \mathfrak{A}_i is \mathfrak{G}_i -central for every $i = 1, \dots, r$. Then \mathfrak{A} is \mathfrak{G} -central.

To prove this result we see that every \mathfrak{G} -ideal of \mathfrak{A} has the form $\mathfrak{B} = \mathfrak{B}_1 + \cdots \oplus \mathfrak{B}_r$ where \mathfrak{B}_i is the ideal of $(\mathfrak{A}_i)_\#$ which is the intersection of \mathfrak{B} and $(\mathfrak{A}_i)_\#$. Then $\mathfrak{B}_i(\mathfrak{H}_i)_\#$ is in \mathfrak{B} and in $(\mathfrak{A}_i)_\#$ and is in \mathfrak{B}_i , \mathfrak{B}_i is an $(\mathfrak{H}_i)_\#$ -ideal. Since \mathfrak{A}_i is central $\mathfrak{B}_i = 0$ or $(\mathfrak{A}_i)_\#$. The remainder of our proof is exactly as in the proof of Theorem 1.

6. Crossed extensions

A quantity g of an algebra \mathfrak{A} has been called *non-singular* if it is neither a right nor a left-divisor of zero. Then the right multiplication R_g and the left multiplication L_g are non-singular and we have

LEMMA 13. *Let a and g be quantities of an algebra \mathfrak{A} such that g is non-singular. Then if an ideal B of A contains either $a \cdot g$ or $g \cdot a$ it contains a .*

For $a \cdot g = aR_g$ is in \mathfrak{B} and so is $(aR_g) \cdot g = a(R_g)^2$. If $a(R_g)^k$ is in \mathfrak{B} so is $[a(R_g)^k] \cdot g = a(R_g)^{k+1}$. Hence \mathfrak{B} contains $a(R_g)^k$ for every positive integer k . Since \mathfrak{B} is a linear space it contains every $a\phi(R_g)$ for $\phi(R_g) = \alpha_1 R_g + \alpha_2 (R_g)^2 + \dots + \alpha_r (R_g)^r$ and the α_i in \mathfrak{F} . But the constant term β_n of the characteristic function $\psi(\lambda) = \lambda^n + \beta_1 \lambda^{n-1} + \dots + \beta_n$ of R_g is not zero, $\psi(R_g) = 0$, the identity transformation $I = -\beta_n^{-1}[\beta_{n-1} R_g + \dots + (R_g)^n]$ is $a\phi(R_g)$, \mathfrak{B} contains $aI = a$. Similarly if \mathfrak{B} contains $g \cdot a$ it contains every $a\phi(L_g)$ and a .

We now make the

DEFINITION. *Let \mathfrak{A} be an algebra with a unity quantity e and \mathfrak{G} be an extending group for \mathfrak{A} . Then a set*

$$(7) \quad \mathfrak{g} = \{g_{s,\tau}\}$$

of quantities $g_{s,\tau}$ of \mathfrak{A} will be called an extending set⁹ for \mathfrak{A} by \mathfrak{G} if g contains one and only one $g_{s,\tau}$ for every pair of transformations S and T of \mathfrak{G} , the $g_{s,\tau}$ are non-singular quantities of \mathfrak{A} , and

$$(8) \quad g_{I,S} = g_{S,I} = e,$$

for every S of \mathfrak{G} .

We shall now proceed to our definition of new classes of algebras. We let \mathfrak{A} be any algebra with a unity quantity e , n be the order over \mathfrak{F} of \mathfrak{A} , \mathfrak{G} be a finite¹⁰ extending group of order m for \mathfrak{A} , \mathfrak{g} be an extending set for \mathfrak{A} by \mathfrak{G} . We let \mathfrak{S} be a subset of \mathfrak{G} containing the identity transformation. Then we shall define an algebra

$$(9) \quad \mathfrak{E} = (\mathfrak{A}, \mathfrak{G}, \mathfrak{S}, \mathfrak{g}),$$

of order $\nu = nm$ over \mathfrak{F} , which we shall call the *crossed extension* of \mathfrak{A} by \mathfrak{S} and \mathfrak{G} with extension set \mathfrak{g} . We shall also call the integer m the *extension index* of \mathfrak{A} under \mathfrak{E} .

We let \mathfrak{N} be a linear space of order ν over \mathfrak{F} so that \mathfrak{N} is the supplementary sum

$$(10) \quad \mathfrak{N}_1 + \dots + \mathfrak{N}_m,$$

⁹ The term extending set is preferable to that of factor set which we reserve for extending sets restricted so that the algebras we construct will be associative.

¹⁰ It seems clear that if we take \mathfrak{G} to be an infinite group our construction will be valid if we take the corresponding linear space \mathfrak{N} to consist of vectors with finitely many non-zero coordinates. Moreover it seems that our hypotheses insuring that the result is a simple algebra will be also sufficient for the algebras of infinite order. It would be interesting to take the case where \mathfrak{G} consists of the non-zero quantities of a division algebra, as well as that where \mathfrak{N} is a Hilbert space.

of linear subspaces \mathfrak{N}_i each of order n over \mathfrak{F} . Then there exist corresponding non-singular linear transformations C_1, \dots, C_m in (\mathfrak{F}) , such that $C_1 = I$, is the identity transformation on \mathfrak{N} ,

$$(11) \quad \mathfrak{N}_i = \mathfrak{N}_1 C_i \quad (i = 1, \dots, m).$$

Thus every quantity of \mathfrak{N} is uniquely expressible in the form

$$(12) \quad a = a_1 C_1 + \dots + a_m C_m \quad (a, \text{ in } \mathfrak{N}_1).$$

Observe next that the linear spaces \mathfrak{A} and \mathfrak{N} have the same order and thus that it is possible to take $\mathfrak{A} = \mathfrak{N}_1$. We do this and have thus imbedded the algebra \mathfrak{A} in \mathfrak{N} as a linear subspace of \mathfrak{N} . We shall actually define \mathfrak{E} to be an algebra whose quantities are the vectors of \mathfrak{N} and we shall formulate our definition so that \mathfrak{A} will be a subalgebra of \mathfrak{E} .

Let us order the transformations of \mathfrak{G} in any order such that the first transformation is I and thus have the notations

$$(13) \quad S = I, S_2, \dots, S_m$$

for these transformations. We have then defined a one-to-one mapping

$$(14) \quad S_i \rightarrow C_i = C_{S_i}$$

of \mathfrak{G} on the set of C_i such that $C_1 = C_I = I$, the identity transformation on the space of order ν . If $S = S_i$ we designate by a_s the coefficient a_i of $C_i = C_s$ and may thus write every a of \mathfrak{N} uniquely in the form

$$(15) \quad a = \sum_s a_s C_s,$$

for the a_s in \mathfrak{A} where the sum is taken over all S of \mathfrak{G} . Write

$$(16) \quad x = \sum_T x_T C_T,$$

and define

$$a \cdot x = \sum_U y_U C_U,$$

for the x_T and y_U in \mathfrak{A} where the y_U are to be determined. Then the distributive law holds only if

$$(17) \quad a \cdot x = \sum_{s, T} (a_s C_s) \cdot (x_T C_T).$$

We now let

$$(18) \quad a_s C_s \cdot x_T C_T = y_{s, T} C_{sT},$$

so that if we write $U = ST$, $T = S^{-1}U$, then we have

$$(19) \quad y_U = \sum_s y_{s, s^{-1}U}.$$

We shall then complete our definition when we express the $y_{s,T}$ in terms of terms of a_s , x_T and g , and we do this by defining the function

$$(20) \quad w(S, T, a, x) = aT \cdot x \quad (S \text{ in } \mathfrak{G}),$$

$$(21) \quad w(S, T, a, x) = x \cdot aT \quad (S \text{ not in } \mathfrak{G}),$$

for every ordered pair a, x of quantities of \mathfrak{A} where the products indicated are *products in* \mathfrak{A} of its quantities. We are then able to write the desired formulas

$$(22) \quad y_{s,T} = w(ST, I, g_{s,T}, w_{s,T}), \quad w_{s,T} = w(S, T, a_s, x_T).$$

In particular $e = g_{I,T} = g_{S,I}$ for every S and hence we have

$$(23) \quad y_{I,T} = w_{I,T} = a_I T \cdot x_T, \quad y_{S,I} = w_{S,I} = w(S, I, a_s, x_I).$$

Conversely let $y_{s,T}$ be defined by (20), (21), (22), so that the y_U are defined uniquely by (19). Since \mathfrak{A} is a linear algebra it is clear that the $y_{s,T}$ are linear in the x_T and thus also in the coordinates of x , the y_U are linear in x , $a \cdot x$ is linear in x . Also every T is a linear transformation, the $y_{s,T}$ are linear in $a_s T$ and hence in a_s , $a \cdot x$ is linear in a . It follows that \mathfrak{E} is a linear algebra.

We note that $y_{s,T} = 0$ if either a_s or x_T is zero. Then if α and κ are in \mathfrak{A} we have $\alpha = \alpha_I, \kappa = x_I$, all the $y_{s,T}$ are zero except $y_{I,I} = \alpha_I \cdot x_I$ by (23), \mathfrak{A} is a subalgebra of \mathfrak{E} . In fact we may prove

THEOREM 7. *The algebra \mathfrak{E} of (7)–(21) contains \mathfrak{A} as a subalgebra and the unity quantity e of \mathfrak{A} is the unity quantity of \mathfrak{E} . Every quantity of \mathfrak{E} is uniquely expressible in the form*

$$(24) \quad \alpha = \sum_s u_s \cdot a_s \quad (S \text{ in } G, a_s \text{ in } \mathfrak{A}),$$

where $u_s = eC_s$, $u_I = e$. The quantities u_s are non-singular quantities of E such that $u_s \cdot a_s = a_s C_s$ and

$$(25) \quad u_s \cdot u_T = u_{sT} g_{s,T},$$

for every S and T of \mathfrak{G} . Then the definitive properties of E are completely given by (20), (21), (25) and

$$(26) \quad (u_s \cdot a_s) \cdot (u_T \cdot x_T) = (u_s \cdot u_T) \cdot [w(S, T, a_s, x_T)].$$

For by (23) we have $y_{I,T} = x_T = y_T$ if $\alpha = e$, $e \cdot \kappa = \kappa$. Similarly $y_{S,I} = w(S, I, a_s, e) = a_s$ if $\kappa = e$, $y_{s,T} = 0$ unless $S^{-1}U = T = I$, $S = U$. Then $y_U = y_{U,I} = a_U$, $y = \alpha = \alpha \cdot e$, e is the unity quantity of \mathfrak{E} . Now $u_s \cdot a_s = eC_s \cdot a_s C_I = y_{s,I} C_s = [w(S, I, e, a_s)]C_s = a_s C_s$ so that (15) is equivalent to (24). Also (18) becomes

$$(27) \quad (u_s \cdot a_s) \cdot (u_T \cdot x_T) = u_{sT} \cdot y_{s,T}.$$

But (25) follows from (27) if we put $a_s = x_T = e$ and use the property that $eT = e$. The definition (22) then states that (27) is equivalent to

$$(28) \quad (u_s \cdot a_s) \cdot (u_T \cdot x_T) = u_{sT} \cdot [w(ST, I, g_{s,T}, w_{s,T})].$$

But $(u_s \cdot u_T) \cdot w_{s,T} = (u_{sT} \cdot g_{sT}) \cdot (u_I \cdot w_{s,T}) = u_{sT} \cdot [w(ST, I, g_{s,T}, w_{s,T})]$ and (28) implies (26). Conversely (26) and (25) imply that $(u_s \cdot a_s) \cdot (u_T \cdot x_T) = (u_{sT} \cdot g_{s,T}) \cdot w_{s,T} = (u_{sT} \cdot g_{s,T}) \cdot (u_I w_{s,T})$ and the fact that $u_{sT} \cdot u_I = u_{sT}$ used with (26) implies (27).

Since $g_{s,T}$ is non-singular we have $u_s \cdot (u_T \cdot x_T) = u_{sT} \cdot y_{s,T}$ where $y_{s,T} = g_{s,T} \cdot x_T$ or $x_T \cdot g_{s,T}$ is not zero unless $x_T = 0$. Then $u_s \cdot \kappa = \sum_T u_{sT} \cdot y_{sT} \neq 0$ unless $\kappa = 0$, u_s is not a left divisor of zero. Similarly u_s is not a right divisor of zero and is a non-singular quantity of \mathfrak{A} . This proves our theorem.

7. A non-simple crossed extension

The crossed extension \mathfrak{E} need not be simple even when \mathfrak{A} is \mathfrak{G} -simple, nor need \mathfrak{E} be central when \mathfrak{A} is \mathfrak{G} -central. Let us give an example here of such an algebra. We take \mathfrak{F} to be a field of real numbers, $\mathfrak{A} = \mathfrak{A}_1 \oplus \mathfrak{A}_2$ where e_1 and e_2 are the respective unity quantities of the quadratic fields \mathfrak{A}_1 and \mathfrak{A}_2 defined by

$$u_1^2 = -e_1, \quad u_2^2 = -e_2, \quad \mathfrak{A}_1 = e_1\mathfrak{F} + u_1\mathfrak{F}, \quad \mathfrak{A}_2 = e_2\mathfrak{F} + u_2\mathfrak{F}.$$

Then every quantity of \mathfrak{A} is uniquely expressible in the form

$$(29) \quad a = \alpha_1 u_1 + \alpha_2 u_2 + \alpha_3 u_3 + \alpha_4 u_4 \quad (\alpha_i \text{ in } \mathfrak{F}),$$

where $u_3 = e_1 - u_1$, $u_4 = e_2 - u_2$. We let \mathfrak{G} be the group of linear transformations on \mathfrak{A} obtained by applying the permutations (13), (24), (13) (24), (12) (34), (14) (23), (1432), (1234) and the identity to the subscripts i on the u_i in (29). Then \mathfrak{G} is an extending group of order eight for \mathfrak{A} and is known¹¹ to be generated by the transformation S obtained by applying (13) and the transformation P obtained by applying (12) (34). We let T be the transformation obtained by applying (24) and have

$$(30) \quad ST = TS, \quad S^2 = T^2 = I, \quad SP = PT, \quad PS = TP.$$

Here TP is the transformation obtained by applying the cycle (1234) to the subscripts of the basal quantities in (29), and PT is its inverse obtained by applying (4321).

The algebra \mathfrak{A} has the property that $\mathfrak{A}_{\mathfrak{R}} = \mathfrak{B}_1 \oplus \mathfrak{B}_2 \oplus \mathfrak{B}_3 \oplus \mathfrak{B}_4$ where \mathfrak{B}_i has order one over the field \mathfrak{R} of all complex numbers, $\mathfrak{B}_i = v_i \mathfrak{F}$ such that $2v_1 = u_3 + (1+i)u_1$, $2v_3 = u_3 + (1-i)u_1$, $2v_2 = u_4 + (1+i)u_2$, $2v_4 = u_4 + (1-i)u_2$ are pairwise orthogonal idempotents. Any non-zero ideal \mathfrak{B} of $\mathfrak{A}_{\mathfrak{R}}$ contains one of the v_i . If \mathfrak{B} is a $\mathfrak{G}_{\mathfrak{R}}$ -ideal it contains with v_1 the quantity u_3 and then we apply P and its powers to get all the u_i , $\mathfrak{B} = \mathfrak{A}_{\mathfrak{R}}$. If \mathfrak{B} contains v_3 and hence $2u_1 + u_3$ we apply S to get $2u_3 + u_1$ and hence the quantity $2(2u_1 + u_3) - (2u_3 + u_1) = 3u_1$. Again $\mathfrak{B} = \mathfrak{A}_{\mathfrak{R}}$. Similarly if \mathfrak{B} contains either v_2 or v_4 it is equal to $\mathfrak{A}_{\mathfrak{R}}$, $\mathfrak{A}_{\mathfrak{R}}$ is \mathfrak{G} -simple, \mathfrak{A} is \mathfrak{G} -central.

Observe that S carries quantities of \mathfrak{A}_1 into other quantities and leaves e_1 as well as every quantity of \mathfrak{A}_2 unaltered, T carries quantities of \mathfrak{A}_2 into other

¹¹ cf. L. E. Dickson, *Modern Algebraic Theories*, pp. 145-6.

quantities such that e_2 is unaltered and T leaves all quantities of \mathfrak{A}_1 unaltered. We form the crossed extension \mathfrak{E} defined above for all quantities in the extension set equal to the unity quantity of \mathfrak{A} and for \mathfrak{G} the identity group. Let $\mathfrak{z} = u_S \cdot e_2 + u_T \cdot e_1$. Now $\mathfrak{z} \cdot a = u_S \cdot (a_2 \cdot e_2) + u_T \cdot (a_1 \cdot e_1)$ for every $a = a_1 + a_2$ such that a_1 is in \mathfrak{A}_1 and a_2 in \mathfrak{A}_2 . Also $a \cdot \mathfrak{z} = u_S \cdot (a_S \cdot e_2) + u_T \cdot (aT \cdot e_1) = \mathfrak{z} \cdot a$ since $(aS) \cdot e_2 = (a_1S + a_2) \cdot e_2 = a_2 \cdot e_2$, $aT \cdot e_1 = (a_2S + a_1) \cdot e_1 = a_1 \cdot e_1$. Moreover $u_S \cdot \mathfrak{z} = u_{ST} \cdot e_1 + e_2 = \mathfrak{z} \cdot u_S = u_{TS} \cdot e_1 + e_2$, $u_T \cdot \mathfrak{z} = u_{TS} \cdot e_2 + e_1 = \mathfrak{z} \cdot u_T$. Finally $u_P \cdot \mathfrak{z} = u_{PS} \cdot e_2 + u_{PT} \cdot e_1$, $\mathfrak{z} \cdot u_P = u_{SP} \cdot e_1 + u_{TP} \cdot e_2 = u_P \cdot \mathfrak{z}$ by (30). That $(a \cdot \mathfrak{x}) \cdot \mathfrak{y} = a \cdot (\mathfrak{x} \cdot \mathfrak{y})$ when one of the factors is \mathfrak{z} follows from the fact that \mathfrak{A} is a commutative associative algebra and that $u_Q \cdot u_R = u_{QR}$. Then \mathfrak{z} is in the center of our crossed extension \mathfrak{E} . But $\mathfrak{z}^2 = (u_S \cdot e_2)^2 + (u_T \cdot e_1)^2 + (u_S \cdot e_2)(u_T \cdot e_1) + (u_T \cdot e_1)(u_S \cdot e_2) = e$ since $(u_S \cdot e_2)^2 = e_2$, $(u_T \cdot e_1)^2 = e_1$, and the other two terms are equal to $u_{ST} \cdot (e_1 \cdot e_2) = 0$. It follows that \mathfrak{E} is a direct sum $\mathfrak{E} = \mathfrak{E}_{\mathfrak{z}_1} \oplus \mathfrak{E}_{\mathfrak{z}_2}$, $2\mathfrak{z}_1 = e - \mathfrak{z}$, $2\mathfrak{z}_2 = e + \mathfrak{z}$, \mathfrak{E} is neither simple nor central.

8. Simple crossed extensions

We have just seen that \mathfrak{A} may be \mathfrak{G} -simple but its extension \mathfrak{E} not a simple algebra. Thus we shall have to make additional hypotheses if we wish every \mathfrak{E} defined for the given \mathfrak{A} , \mathfrak{G} , \mathfrak{S} to be simple. These conditions are really a part of the usual associative crossed product definition, but are hidden in the more explicit and special nature of those algebras.

Let us call a linear transformation S on the linear space \mathfrak{A} an *inner*¹² or an *outer* transformation for this algebra according as there is or is not a quantity $b \neq 0$ in \mathfrak{A} such that

$$(31) \quad b \cdot x = xS \cdot b.$$

The identity transformation I is one of a class of inner transformations on \mathfrak{A} which have the property that (31) holds for b in the center of \mathfrak{A} . We shall call any such transformation a *semi-identity* transformation for \mathfrak{A} . If S is semi-identical and \mathfrak{A} is semi-simple we may write $b = b_1 + \cdots + b_s$, $\mathfrak{A} = \mathfrak{A}_1 \oplus \cdots \oplus \mathfrak{A}_s \oplus C$ for simple algebras \mathfrak{A}_i such that $b_i \neq 0$ is in the center of \mathfrak{A}_i and (31) becomes $b \cdot (x - xS) = 0$. But there exist d_1, \dots, d_s in the center of $\mathfrak{A}_1, \dots, \mathfrak{A}_s$ such that $d_i \cdot b_i = e_i$, $f = e_1 + \cdots + e_s$ is the unity quantity of the ideal $\mathfrak{B} = \mathfrak{A}_1 \oplus \cdots \oplus \mathfrak{A}_s$ of \mathfrak{A} , $f \cdot (x - xS) = 0$. Then $\mathfrak{A} = \mathfrak{B} \oplus \mathfrak{C}$ such that $y - yS$ is in \mathfrak{C} for every y of \mathfrak{B} , $\mathfrak{C}S = \mathfrak{C}$, $b \cdot x = xS \cdot b$ for every b in the center of \mathfrak{B} .

We now have a terminology for the hypotheses we shall require, and we shall prove

THEOREM 8. *Let \mathfrak{A} be a semi-simple algebra, \mathfrak{G} be an extending group for \mathfrak{A} such that \mathfrak{A} is \mathfrak{G} -simple and I is the only semi-identity transformation for \mathfrak{A} in \mathfrak{G} .*

¹² If \mathfrak{S} is an automorphism it is inner in the ordinary sense only when the quantity b is non-singular. However we shall require the (only slightly) more general hypothesis we give here.

Then if there is any subset \mathfrak{S} of \mathfrak{G} such that \mathfrak{S} consists of I and outer¹³ transformations for \mathfrak{A} the crossed extensions $\mathfrak{E} = (\mathfrak{A}, \mathfrak{G}, \mathfrak{S}, g)$ are simple algebras.

For let $\rho(a)$ be the number of non-zero coefficients a_s in the unique expression (24) of any a in \mathfrak{E} for a_s in \mathfrak{A} . Then $\rho(0) = 0$, $\rho(a)$ is a positive integer for every $a \neq 0$. Let \mathfrak{B} be a non-zero ideal of \mathfrak{E} and ρ be the least $\rho(b)$ for any $b \neq 0$ in \mathfrak{B} . Then some b in \mathfrak{B} has the property $\rho(b) = \rho$ and if we write $b = \sum u_s \cdot b_s$ as in (24) there is an S_0 such that $b_{S_0} \neq 0$. Then $u_T \cdot b = \sum_s u_{Ts} \cdot c_{Ts}$ where c_{Ts} is the product of b_s by a non-singular quantity $g_{T,s}$ of \mathfrak{A} and is not zero if $b_s \neq 0$. Take $T = S_0^{-1}$ and have a quantity c in \mathfrak{B} such that $\rho(c) = \rho$, $c = \sum_s u_s \cdot c_s$, $c_I \neq 0$. We now let \mathfrak{D} be the set of all finite sums of terms of the form $(x \cdot c) \cdot y$, $x \cdot (y \cdot c)$, for x and y in \mathfrak{A} . It follows that every d of \mathfrak{D} has the form

$$d = d_I + \sum_{i=2}^p u_{S_i} d_{S_i},$$

for a fixed set I, S_2, \dots, S_p in \mathfrak{G} and with the d_s in \mathfrak{A} . Then c is in the set \mathfrak{D} and \mathfrak{D} is a non-zero linear subspace of the ideal \mathfrak{B} such that $\rho(b) = \rho$ for every non-zero b of \mathfrak{D} . Moreover the quantities d_I consist of all finite sums of the form $(x \cdot c_I) \cdot y$ or $x \cdot (c_I \cdot y)$ for x and y in \mathfrak{A} , the set \mathfrak{N} of all the d_I is a non-zero ideal of \mathfrak{A} . Let f_I be its unity quantity so that f_I is in the center of \mathfrak{A} and there is a quantity f in \mathfrak{D} such that

$$f = f_I + \sum_{i=2}^p u_{S_i} f_{S_i} \quad (f_{S_i} \text{ in } \mathfrak{A}).$$

If $\rho < 1$ and some $T = S_i$ is not in \mathfrak{S} it is not a semi-identical transformation and there exists a quantity x in \mathfrak{A} such that $h_I = x \cdot f_I - f_I \cdot xT = f_I \cdot (x - xT) \neq 0$. The corresponding $h = x \cdot f - f \cdot xT \neq 0$ is in \mathfrak{D} so that $\rho(h) = \rho$. But the term of h involving u_T is $x \cdot (u_T \cdot f_T) - (u_T \cdot f_T) \cdot xT = u_T \cdot (xT \cdot f_T - xT \cdot f_T) = 0$, a contradiction. Hence every S_i is in \mathfrak{S} and if $\rho > 1$ there is an $S_i = T$ which is an outer transformation, there must exist a quantity x in \mathfrak{A} such that $h_T = xT \cdot f_T - f_T \cdot x \neq 0$, $x \cdot (u_T \cdot f_T) - (u_T \cdot f_T) \cdot x = u_T \cdot h_T \neq 0$ is the term involving u_T in $h = x \cdot f - f \cdot x$. Then $h \neq 0$ is in \mathfrak{D} and $\rho(h) = \rho$. However f_I is in the center of \mathfrak{A} and $h_I = x \cdot f_I - f_I \cdot x = 0$, a contradiction.

This proves that $\rho = 1$ and that the intersection \mathfrak{U} of \mathfrak{B} and \mathfrak{A} is a non-zero ideal of \mathfrak{A} . If \mathfrak{U} contains b and S is in \mathfrak{G} we write $T = S^{-1}$ and have $u_T \cdot (b \cdot u_s) = h$ where $h = g_{T,s} \cdot bS$ or $bS \cdot g_{T,s}$ is in \mathfrak{U} . By Lemma 13 so is dS . Thus \mathfrak{U} is a \mathfrak{G} -ideal of \mathfrak{A} and $\mathfrak{U} = \mathfrak{A}$ since \mathfrak{A} is \mathfrak{G} -simple. Then \mathfrak{B} contains e and is the unit ideal \mathfrak{E} , \mathfrak{E} is a simple algebra.

We note that the example in Section 6 of a crossed extension which is not a simple algebra failed to satisfy our hypotheses precisely in that \mathfrak{G} contained semi-identical transformations $S \neq I$.

We shall not try to compute the center of a crossed extension \mathfrak{E} and so to

¹³ In the case of ordinary crossed products $\mathfrak{G} = \mathfrak{S}$ and \mathfrak{A} is a field. Our hypotheses are then satisfied.

prove that a given \mathfrak{E} is central simple but we shall rather try to see that our hypotheses are in a form such that they hold also when the field \mathfrak{F} is extended.¹⁴ Thus we prove

LEMMA 14. *A linear transformation S on a separable algebra \mathfrak{A} is a semi-identical transformation or an inner transformation for \mathfrak{A} if and only if $S_{\mathfrak{R}}$ has the corresponding property for $\mathfrak{A}_{\mathfrak{R}}$, where \mathfrak{R} is any scalar extension of \mathfrak{F} .*

For if (31) holds for a quantity b in \mathfrak{A} and every x of \mathfrak{A} we will also have $b \cdot x = x S_{\mathfrak{R}} \cdot b$ for every x in $\mathfrak{A}_{\mathfrak{R}}$, $S_{\mathfrak{R}}$ is inner when S is. Also $S_{\mathfrak{R}}$ is semi-identical when S is since if \mathfrak{Z} is the center of \mathfrak{A} the center of $\mathfrak{A}_{\mathfrak{R}}$ is $\mathfrak{Z}_{\mathfrak{R}}$. Conversely let $y \cdot x = x S_{\mathfrak{R}} \cdot y$ for every x of $\mathfrak{A}_{\mathfrak{R}}$ and a fixed quantity y in $\mathfrak{A}_{\mathfrak{R}}$. Then we may write $y = y_1 \xi_1 + \cdots + y_s \xi_s$ for y_j in \mathfrak{A} where the ξ_j are in \mathfrak{R} and are such that $a_1 \xi_1 + \cdots + a_s \xi_s = 0$ for a_i in \mathfrak{A} if and only if the a_i are all zero. Take x in \mathfrak{A} and so obtain $x S_{\mathfrak{R}} = x S$ in \mathfrak{A} , $y \cdot x - x S_{\mathfrak{R}} \cdot y = \sum (y_j \cdot x - x S_{\mathfrak{R}} \cdot y_j) \xi_j = 0$, $y_j \cdot x = x S \cdot y_j$. Hence if $S_{\mathfrak{R}}$ is inner so is S . If $S_{\mathfrak{R}}$ is semi-identical the quantity y may be taken to be in $\mathfrak{Z}_{\mathfrak{R}}$, y_1 is in \mathfrak{Z} , S is semi-identical.

We may thus apply Lemma 14 to Theorem 8 and obtain

THEOREM 9. *Let \mathfrak{A} be a separable algebra,¹⁵ \mathfrak{G} be an extending group for \mathfrak{A} such that \mathfrak{A} is \mathfrak{G} -central and I is the only semi-identity transformation for \mathfrak{A} in \mathfrak{G} . Then if \mathfrak{S} is any subset of \mathfrak{G} consisting of I and outer transformations for \mathfrak{A} the crossed extensions $\mathfrak{E} = (\mathfrak{A}, \mathfrak{G}, \mathfrak{S}, \mathfrak{g})$ are central simple algebras.*

If \mathfrak{A} is a simple algebra the only semi-identity transformation for \mathfrak{A} is I and we have

THEOREM 10. *Let \mathfrak{G} be an extending group for a simple algebra \mathfrak{A} , \mathfrak{S} consist of I and outer transformations for \mathfrak{A} in \mathfrak{G} . Then every crossed extension of \mathfrak{A} is a simple algebra.*

If \mathfrak{S} consists of I alone we shall write

$$\mathfrak{E} = (\mathfrak{A}, \mathfrak{G}, \mathfrak{g})$$

for the corresponding crossed extensions. These are surely the most interesting of our new algebras and we shall state the results of Theorems 8 and 9 for such algebras as

THEOREM 11. *Let \mathfrak{G} be an extending group for a simple algebra \mathfrak{A} . Then every crossed extension $\mathfrak{E} = (\mathfrak{A}, \mathfrak{G}, \mathfrak{g})$ is a simple algebra. Moreover if \mathfrak{A} is central simple so is \mathfrak{E} .*

9. Associativity

A crossed extension \mathfrak{E} is associative if and only if \mathfrak{A} is associative and $[(u_s \cdot a_s)(u_T \cdot x_T)] \cdot u_P \cdot w_P = (u_s \cdot a_s) \cdot [(u_T \cdot x_T) \cdot (u_P \cdot w_P)]$ for every a_s, x_T, w_P of \mathfrak{A} and S, T, P of \mathfrak{G} . If $\mathfrak{G} \neq \mathfrak{S}$ we take S not in \mathfrak{S} , $a_s = e$, $T = P = I$ and have $u_s \cdot (x \cdot w) = (u_s \cdot x) \cdot w = u_s(w \cdot x)$ which is possible in an associative

¹⁴ This seems to be the best possible method of proof even for ordinary crossed products.

¹⁵ In the associative case the simple components of \mathfrak{A} are necessarily equivalent, since \mathfrak{A} is \mathfrak{G} -simple and $\mathfrak{G} = \mathfrak{S}$ is composed of automorphisms. However this does not appear to be necessary here and this question should be studied.

algebra \mathfrak{E} if and only if \mathfrak{A} is commutative. But when \mathfrak{A} is commutative the algebras $\mathfrak{E} = (\mathfrak{A}, \mathfrak{G}, \mathfrak{S}, g)$ are the same for every \mathfrak{S} and we may take $\mathfrak{E} = (\mathfrak{A}, \mathfrak{G}, g)$. Hence we have $\mathfrak{S} = [I]$ in every case. We now compute $(u_s \cdot a_s) \cdot (u_T \cdot x_T) = u_{sT} \cdot (g_{s,T} \cdot a_s T \cdot x_T)$. Similarly $(u_T \cdot x_T) \cdot (u_P \cdot w_P) = u_{TP} \cdot (g_{T,P} \cdot x_T P \cdot w_P)$. Multiply the first of these products on the right by $u_P \cdot w_P$ and the second on the left by $u_s \cdot a_s$. The resulting products each have the non-singular left factor u_{sTP} , and if \mathfrak{E} is associative we have

$$(32) \quad g_{sT,P} [g_{s,T} \cdot (a_s T \cdot x_T)] P \cdot w_P = g_{s,TP} \cdot (a_s TP) \cdot [g_{T,P} \cdot (x_T P \cdot w_P)].$$

Put $a_s = x_T = w_P$ equal to the unity quantity e of \mathfrak{A} and obtain

$$(33) \quad g_{sT,P} \cdot (g_{s,T} P) = g_{s,TP} g_{T,P} \quad (S, T, P \text{ in } \mathfrak{G}).$$

Next put $S = T = I$ and $w_P = e$ and so obtain

$$(34) \quad (a \cdot x) P = a P \cdot x P,$$

that is, \mathfrak{G} is a group of automorphisms of \mathfrak{A} . Put $x_T = w_P = e$, $P = I$, to see that the $g_{s,T}$ are in the center of \mathfrak{A} . Conversely if (33) holds for an automorphism group \mathfrak{G} and the $g_{s,T}$ in the center of \mathfrak{A} we have (32) and \mathfrak{E} is associative.

We shall call an extension set satisfying (33) for an automorphism group \mathfrak{G} of \mathfrak{A} a *factor set*. Then we have proved

THEOREM 12. *A crossed extension $\mathfrak{E} = (\mathfrak{A}, \mathfrak{G}, \mathfrak{S}, g)$ is associative if and only if \mathfrak{A} is associative, \mathfrak{G} is a group of automorphisms of \mathfrak{A} , g is a factor set of \mathfrak{A} whose quantities are in the center of \mathfrak{A} , and $\mathfrak{G} = \mathfrak{S}$.*

We now have the consequent

COROLLARY I. *Let $\mathfrak{E} = (\mathfrak{A}, \mathfrak{G}, g)$. Then if \mathfrak{A} is not commutative the algebra \mathfrak{E} is not associative.*

COROLLARY II. *Let $\mathfrak{E} = (\mathfrak{A}, \mathfrak{G}, g)$ where \mathfrak{A} is a central simple algebra. Then \mathfrak{E} is a non-associative central simple algebra.*

We have tacitly assumed in all of our work that the order n of \mathfrak{A} is not one and that the order m of \mathfrak{G} is also not one. Then \mathfrak{E} has order nm and \mathfrak{A} is a proper subalgebra of \mathfrak{E} .

10. Explicit construction

Let us indicate some of the types of algebras included under our definition. The first of these are the ordinary crossed products and the special case of cyclic algebras. These are given by $\mathfrak{E} = (\mathfrak{A}, \mathfrak{G}, \mathfrak{S}, g)$ for $\mathfrak{G} = \mathfrak{S}$ the automorphism group of a normal field, g a factor set. Then our construction Theorem 9 implies that \mathfrak{E} is central simple and this seems to be a proof of that result which has been overlooked in the literature. The other algebras of Theorem 12 have been considered only in a rather special case.

Let us restrict all further attention to the case where \mathfrak{A} is central simple and \mathfrak{S} is the identity group so that $\mathfrak{E} = (\mathfrak{A}, \mathfrak{G}, g)$ is not associative and is central simple. Then we shall define one very interesting type of algebra which is the

crossed extension of \mathfrak{A} by what is, essentially, a permutation group. We shall call such algebras *permutation algebras* and define them as follows. We let e, u_2, \dots, u_n be any basis of \mathfrak{A} over \mathfrak{F} and let u_1 be determined by

$$(35) \quad u_1 + \dots + u_n = e.$$

Then u_1, \dots, u_n are a basis of \mathfrak{A} over \mathfrak{F} and we have a unique expression

$$(36) \quad a = \alpha_1 u_1 + \dots + \alpha_n u_n \quad (\alpha_i \text{ in } \mathfrak{F})$$

for every a of \mathfrak{A} . We define

$$(37) \quad aS(P) = \alpha_1 u_{i_1} + \dots + \alpha_n u_{i_n}$$

for every permutation

$$(38) \quad P = \begin{pmatrix} 1 & 2 & \dots & n \\ i_1 & i_2 & \dots & i_n \end{pmatrix}$$

and thus have defined a group \mathfrak{G} of non-singular linear transformations $S = S(P)$ on \mathfrak{A} such that $eS = e$ for every permutation group \mathfrak{G}_0 of permutations P . Clearly \mathfrak{G} is equivalent to \mathfrak{G}_0 and the algebra $\mathfrak{E} = (\mathfrak{A}, \mathfrak{G}, \mathfrak{g})$ is central simple for every \mathfrak{g} . Moreover, this type of algebra is special since, while every finite group \mathfrak{G} may be represented as a permutation group, \mathfrak{G} may not¹⁶ permute any set of basal quantities of A .

Let us give an iterative process next for the construction of a family of central simple algebras defined for what is essentially a single group. We let $\mathfrak{E} = (\mathfrak{A}, \mathfrak{G}, \mathfrak{F}, \mathfrak{g})$ be a given central simple crossed extension of order nm over \mathfrak{F} defined for an algebra \mathfrak{A} of order n and a group \mathfrak{G} of order m . Then every quantity κ of \mathfrak{E} is uniquely expressible in the form

$$\kappa = u_1 \cdot a_1 + \dots + u_m \cdot a_m$$

for the a_i in \mathfrak{A} and $u_i = u_{S_i}$, $S_1 = I$, S_2, \dots, S_m the transformations of \mathfrak{G} . Since \mathfrak{E} is central simple any

$$\mathfrak{E}_0 = (\mathfrak{E}, \mathfrak{G}_0, \mathfrak{g}_0)$$

will be central simple for any extending group \mathfrak{G}_0 and extension set \mathfrak{g}_0 . We let \mathfrak{G}_0 be the set of linear transformations

$$S_0 : \quad \kappa \rightarrow \kappa S_0 = u_1 \cdot a_1 S + \dots + u_m \cdot a_m S \quad (S \text{ in } \mathfrak{G}).$$

Then \mathfrak{G}_0 is a finite group equivalent to \mathfrak{G} and is clearly an extending group for \mathfrak{E} , the algebra $\mathfrak{E}_0 = (\mathfrak{E}, \mathfrak{G}_0, \mathfrak{g}_0)$ is central simple for every \mathfrak{g}_0 and we may

¹⁶ It would be desirable now to prove the existence of examples of a simple algebra \mathfrak{A} and an extending group \mathfrak{G} such that no basis of \mathfrak{A} exists for which \mathfrak{G} may be regarded as a permutation group. Observe also that for every field \mathfrak{A} of order n over \mathfrak{F} we have a simple permutation algebra. This is then a generalization of the crossed product concept where \mathfrak{A} is a normal field, the crossed product is a permutation algebra defined for a normal basis of \mathfrak{A} .

indeed choose g_0 in \mathfrak{A} . This process may be repeated to obtain central simple algebras of order nm^t for every \mathfrak{G} of order m .

In particular we have the *generalized crossed products*

$$\mathfrak{E}_t = (\mathfrak{N}, \mathfrak{G}, g_1, \dots, g_t),$$

where \mathfrak{N} is a normal field of order r , \mathfrak{G} is its automorphism group, the g_i are all factor sets (or merely any extension sets). This algebra has order r^t over \mathfrak{F} . We thus have the *generalized cyclic algebras*

$$(\mathfrak{N}, S, \gamma_1, \dots, \gamma_t)$$

for the $\gamma_i \neq 0$ in \mathfrak{F} .

Another process of iteration is that where we define S_0 by $\ast S_0 = u_1 \cdot aS + u_2 \cdot a_2 + \dots + u_m \cdot a_m$ and there are other obvious variations. However these may possibly give corresponding algebras obtained from the type given above by the use of a different \mathfrak{G} and extension set. Thus we are led to the problem of determining when two crossed extensions defined for the same \mathfrak{A} but with distinct groups and extension sets are equivalent, and also when they are isotopic. The solution of this problem will require a study of automorphisms of our algebras and also a solution of the simpler problem of determining conditions that $(\mathfrak{A}, \mathfrak{G}, \mathfrak{S}, g)$ shall equal $(\mathfrak{A}, \mathfrak{G}, \mathfrak{S}, f)$ for given distinct extension sets g and f .

The associative algebra theory suggests for study many other fundamental problems regarding our new classes of algebras. For example let us call any algebra equivalent to a generalized cyclic algebra $(\mathfrak{N}, S, 1, \dots, 1)$ a *generalized total matric algebra* over \mathfrak{F} . Then we seek to study the nature of the simple subalgebras of all such algebras (as well as of all other crossed extensions) and in particular to prove that they are all generalized cyclic algebras if \mathfrak{F} is a p -adic or an algebraic number field. Such a study would probably require a study of splitting fields, direct products, the \mathfrak{G} -centralizer of a simple subalgebra of \mathfrak{E} , further extension of the concept of total matric algebra, similarity for crossed extensions, a theory of division algebras and a theory of exponents. It seems clear that our crossed extension definition includes many new varieties of simple algebras and it should lead to a host of new applications of modern algebraic techniques.

EXTENSIONS OF DIFFERENTIAL FIELDS, I

By E. R. KOLCHIN¹

(Received May 6, 1942)

Introduction

It is a well-known theorem of algebra that a finite algebraic extension of a field of characteristic zero K always contains a primitive element ω :

$$K(\alpha_1, \dots, \alpha_n) = K(\omega).$$

Moreover, by means of the theory of Galois, it is possible to characterize those elements of the extension which are primitive.² The present paper treats the analogous problems for differential fields (ordinary or partial).

A simple example shows that the precise analog is not true without further restriction. Let \mathfrak{F}_0 be the ordinary differential field of rational numbers, and let α_1 and α_2 be two algebraically independent complex constants. Since α_1 and α_2 both have zero derivatives, $\mathfrak{F}_0\langle\alpha_1, \alpha_2\rangle$ is set-theoretically identical with $\mathfrak{F}_0(\alpha_1, \alpha_2)$,³ whence it is clear that there exists no number $\beta \in \mathfrak{F}_0\langle\alpha_1, \alpha_2\rangle$ such that $\mathfrak{F}_0\langle\alpha_1, \alpha_2\rangle = \mathfrak{F}_0\langle\beta\rangle$. However, for the theorem in question to hold it suffices to place a mild condition on the differential field. In the ordinary case the condition reduces to the requirement that *the differential field contain a non-constant* (that is, an element whose derivative is different from zero), in the general (partial) case, the condition is that *the differential field contain a set of elements whose Jacobian does not vanish*.

In studying those elements of an extension \mathfrak{G} of a differential field \mathfrak{F} which are primitive, a theorem presents itself which bears a similarity to results from Galois' theory. However any attempt in this direction seems destined to but fragmentary results, as the concept analogous to a *normal* extension of a field is lacking, so that one must speak of isomorphisms instead of automorphisms, thereby abandoning the concept of group.

1. Generic solutions

Throughout this paper \mathfrak{F} will denote a differential field of characteristic zero with m types of differentiation $\delta_1, \dots, \delta_m$,⁴ and y_1, \dots, y_n will denote unknowns (m and n are positive integers).

Let Σ be a system of differential polynomials in $\mathfrak{F}\{y_1, \dots, y_n\}$ with mani-

¹ National Research Fellow.

² This is not, of course, the simplest characterization.

³ $\mathfrak{F}\langle u, \dots \rangle$ means the result of the differential field adjunction to \mathfrak{F} of the elements u, \dots . $\mathfrak{F}(u, \dots)$ means, as usual, the result of the field adjunction to \mathfrak{F} (considered as a field) of the elements u, \dots . The result of differential ring adjunction is indicated by curled brackets: $\mathfrak{F}\{u, \dots\}$.

⁴ This concept has been discussed by H. W. Raudenbush, Bulletin of the American Mathematical Society, vol. 40 (1934), pp. 714-720.

fold \mathfrak{M} . A set η_1, \dots, η_n of elements of a differential extension field of \mathfrak{F} will be called a *generic solution of Σ* (or of \mathfrak{M} , with respect to \mathfrak{F}) if a necessary and sufficient condition for a differential polynomial $F(y_1, \dots, y_n)$ in $\mathfrak{F}\{y_1, \dots, y_n\}$ to belong to Σ is

$$F(\eta_1, \dots, \eta_n) = 0.$$

It is easy to see that if Σ has a generic solution, then Σ is a prime differential ideal in $\mathfrak{F}\{y_1, \dots, y_n\}$, so that \mathfrak{M} is irreducible over \mathfrak{F} . Conversely if Σ is a prime differential ideal other than the whole ring $\mathfrak{F}\{y_1, \dots, y_n\}$, then Σ has a generic solution. For example, if, in the differential ring of remainder classes $\mathfrak{F}\{y_1, \dots, y_n\}/\Sigma$, \bar{y}_i is the remainder class containing y_i , then $\bar{y}_1, \dots, \bar{y}_n$ are elements of a differential field containing \mathfrak{F} (namely, the differential field of quotients of $\mathfrak{F}\{y_1, \dots, y_n\}/\Sigma$), and $F(y_1, \dots, y_n)$ is in Σ if and only if $F(\bar{y}_1, \dots, \bar{y}_n) = 0$. It is not hard to see moreover, that any generic solution η_1, \dots, η_n of Σ is equivalent to $\bar{y}_1, \dots, \bar{y}_n$, that is, $\eta_i \rightarrow \bar{y}_i$ ($i = 1, \dots, n$) generates an isomorphism:

$$\mathfrak{F}\langle \eta_1, \dots, \eta_n \rangle \cong \mathfrak{F}\langle \bar{y}_1, \dots, \bar{y}_n \rangle.^5$$

Now, a prime differential ideal Σ in $\mathfrak{F}\{y_1, \dots, y_n\}$ may very well decompose, over an extension \mathfrak{G} of \mathfrak{F} , into several essential prime differential ideals:

$$(1) \quad \{\Sigma\} = \Lambda_1 \cap \dots \cap \Lambda_s, \quad \text{in } \mathfrak{G}\{y_1, \dots, y_n\}.$$

Let ζ_1, \dots, ζ_n be a generic solution of some Λ_i , say of Λ_h . Then ζ_1, \dots, ζ_n is a generic solution of Σ . Indeed, it is clear that $F(y_1, \dots, y_n) \in \Sigma$ implies $F(\zeta_1, \dots, \zeta_n) = 0$, as $\Sigma \subseteq \Lambda_h$. Conversely, suppose that $F = F(y_1, \dots, y_n) \in \mathfrak{F}\{y_1, \dots, y_n\}$, and that $F(\zeta_1, \dots, \zeta_n) = 0$. Let

$$G \in \Lambda_1 \cap \dots \cap \Lambda_{h-1} \cap \Lambda_{h+1} \cap \dots \cap \Lambda_s, \quad G \notin \Lambda_h.$$

Then FG vanishes for all solutions of Σ , so that, by the Ritt analog of the *Nullstellensatz*, some power $(FG)^k$ is a linear combination, with coefficients in $\mathfrak{G}\{y_1, \dots, y_n\}$, of differential polynomials in Σ :

$$F^k G^k = C_1 S_1 + \dots + C_l S_l \quad (S_i \in \Sigma).$$

The coefficients of G^k , C_1, \dots, C_l are in \mathfrak{G} . Letting $\omega_1, \dots, \omega_\theta$ be, with respect to \mathfrak{F} , a linearly independent linear basis of these coefficients, we find a relation

$$F^k(H_1\omega_1 + \dots + H_\theta\omega_\theta) = T_1\omega_1 + \dots + T_\theta\omega_\theta,$$

where each $T_i \in \Sigma$, each $H_i \in \mathfrak{F}\{y_1, \dots, y_n\}$, and $H_1\omega_1 + \dots + H_\theta\omega_\theta = G^k$. Equating coefficients, on both sides, of the linearly independent elements ω_i , we see that

⁵ The isomorphism indicated by the symbol \cong maps not only the sum and product of two elements onto the sum and product, respectively, of their images, but also the various derivatives of an element onto the corresponding derivatives of its image.

$$F^*H_i = T_i \in \Sigma \quad (i = 1, \dots, g).$$

But not every H_i is in Σ , for otherwise G would be in Λ_A . Hence, since Σ is a prime ideal, $F \in \Sigma$.

We use this result to prove that if the prime differential ideal Σ in $\mathfrak{F}\{y_1, \dots, y_n\}$ has a generic solution η_1, \dots, η_n , if \mathfrak{G} is a differential extension field of $\mathfrak{F}\langle\eta_1, \dots, \eta_n\rangle$, and if no extension of \mathfrak{G} contains another generic solution of Σ , then each $\eta_i \in \mathfrak{F}$.

For, let (1) be the decomposition of $\{\Sigma\}$ into essential prime differential ideals in $\mathfrak{G}\{y_1, \dots, y_n\}$. Any generic solution of Λ_1 is a generic solution of Σ and therefore is identical with η_1, \dots, η_n . The same holds for every Λ_i , so that $s. = 1, \{\Sigma\} = [y_1 - \eta_1, \dots, y_n - \eta_n]$, and the only solution of Σ is η_1, \dots, η_n . Assume, now, that $\eta_1 \notin \mathfrak{F}$. For some k ,

$$(y_1 - \eta_1)^k = C_1 S_1 + \dots + C_l S_l \quad (S_i \in \Sigma).$$

We suppose that k has been chosen as low as possible, so that $1, \eta_1, \dots, \eta_1^k$ are linearly independent over \mathfrak{F} . Letting $1, \eta_1, \dots, \eta_1^k, \omega_1, \dots, \omega_s$ be a linearly independent linear basis, with respect to \mathfrak{F} , of the coefficients in $(y_1 - \eta_1)^k, C_1, \dots, C_l$, and equating coefficients of η_1^k , we arrive at the contradiction that $1 \in \Sigma$. Hence $\eta_1 \in \mathfrak{F}$, similarly, every $\eta_i \in \mathfrak{F}$.

2. Relative isomorphisms

Let \mathfrak{G} be a differential extension field of \mathfrak{F} . By an *isomorphism of \mathfrak{G} with respect to \mathfrak{F}* we shall mean an isomorphic mapping of \mathfrak{G} onto a differential field \mathfrak{G}' such that

- (a) \mathfrak{G}' is an extension of \mathfrak{F} ,
- (b) the isomorphic mapping leaves each element of \mathfrak{F} invariant,
- (c) \mathfrak{G} and \mathfrak{G}' have a common extension.

By means of well-ordering methods it is easy to show that an isomorphism of \mathfrak{G} with respect to \mathfrak{F} can be extended to an automorphism of the common extension of \mathfrak{G} and its map under the isomorphism.

Concerning such relative isomorphisms we prove the following theorem:

Let \mathfrak{G} be an extension of \mathfrak{F} , and let $\gamma \in \mathfrak{G}$. A necessary and sufficient condition that $\gamma \in \mathfrak{F}$ is that every isomorphism of \mathfrak{G} with respect to \mathfrak{F} leaves γ invariant. A necessary and sufficient condition that γ be a primitive element, that is, that $\mathfrak{G} = \mathfrak{F}\langle\gamma\rangle$, is that no isomorphism of \mathfrak{G} with respect to \mathfrak{F} other than the identity leaves γ invariant.

PROOF: A. If $\gamma \in \mathfrak{F}$, then by condition (b), every isomorphism of \mathfrak{G} with respect to \mathfrak{F} leaves γ invariant. Now let $\gamma \notin \mathfrak{F}$, and denote by Γ the prime differential ideal of all differential polynomials in $\mathfrak{F}\{y\}$ which vanish for $y = \gamma$. γ is a generic solution of Γ . Since $\gamma \notin \mathfrak{F}$, we know by §1 that there exists a differential field $\mathfrak{G} \supseteq \mathfrak{G}$ in which Γ has another generic solution γ' . Now, $\gamma \rightarrow \gamma'$ generates an isomorphism between $\mathfrak{F}\langle\gamma\rangle$ and $\mathfrak{F}\langle\gamma'\rangle$ which leaves invariant every element of \mathfrak{F} . This isomorphism can be extended to an automorphism

of \mathfrak{G} , which automorphism in turn can be contracted to produce an isomorphism of \mathfrak{G} with respect to \mathfrak{F} which does not leave γ invariant.

B. If $\mathfrak{G} = \mathfrak{F}\langle\gamma\rangle$, every element of \mathfrak{G} is a rational function, with coefficients in \mathfrak{F} , of γ and its various derivatives, so that an isomorphism of \mathfrak{G} with respect to \mathfrak{F} which leaves γ invariant leaves every element of \mathfrak{G} invariant, that is, is the identity isomorphism. Conversely, if $\mathfrak{G} \neq \mathfrak{F}\langle\gamma\rangle$, there is an element $\alpha \in \mathfrak{G}$ such that $\alpha \notin \mathfrak{F}\langle\gamma\rangle$. By the part of the theorem already proved there is an isomorphism of \mathfrak{G} with respect to $\mathfrak{F}\langle\gamma\rangle$ which does not leave α invariant. This is an isomorphism of \mathfrak{G} with respect to \mathfrak{F} , other than the identity, which leaves γ invariant.

The existence, in certain general cases, of a primitive element will be demonstrated in §4, after the proof of a preparatory result in §3.

3. Non-vanishing of nonzero differential polynomials

The following lemma will be used in §4.

A necessary and sufficient condition that, for an arbitrary nonzero differential polynomial $A = A(y_1, \dots, y_n) \in \mathfrak{F}\{y_1, \dots, y_n\}$, there exist elements $\eta_1, \dots, \eta_n \in \mathfrak{F}$ such that $A(\eta_1, \dots, \eta_n) \neq 0$, is that \mathfrak{F} contain m elements ξ_1, \dots, ξ_m whose Jacobian is different from zero:

$$J = \begin{vmatrix} \delta_1 \xi_1 & \cdots & \delta_m \xi_1 \\ \vdots & \cdots & \vdots \\ \delta_1 \xi_m & \cdots & \delta_m \xi_m \end{vmatrix} \neq 0.$$

PROOF: Necessity. If \mathfrak{F} has the property in question, then, in particular, there are elements ξ_1, \dots, ξ_m which do not annul

$$J(y_1, \dots, y_m) = \begin{vmatrix} \delta_1 y_1 & \cdots & \delta_m y_1 \\ \vdots & \cdots & \vdots \\ \delta_1 y_m & \cdots & \delta_m y_m \end{vmatrix}.$$

Sufficiency. It obviously suffices to consider the case $n = 1$: $A = A(y) \in \mathfrak{F}\{y\}$. Now, since $J \neq 0$ there exists an $m \times m$ matrix (α_{ij}) , with elements in \mathfrak{F} , such that $(\alpha_{ij})(\delta_j \xi_k)$ is the unit matrix. Hence, if we introduce the operators

$$\delta'_i = \alpha_{i1}\delta_1 + \cdots + \alpha_{im}\delta_m \quad (i = 1, \dots, m)$$

in terms of which, in turn, the operators δ_j may be expressed

$$\delta_j = \beta_{j1}\delta'_1 + \cdots + \beta_{jm}\delta'_m \quad (j = 1, \dots, m),$$

we shall have

$$\delta'_i \xi_k = \begin{cases} 1 & \text{if } i = k, \\ 0 & \text{if } i \neq k. \end{cases}$$

Moreover, since

$$\begin{aligned}\delta'_p \delta'_q &= \sum_i \alpha_{pi} \delta_i \sum_j \alpha_{qj} \delta_j \\ &= \sum_i \sum_j \alpha_{pi} \alpha_{qj} \delta_i \delta_j + \sum_j \left(\sum_i \alpha_{pi} \delta_i \alpha_{qj} \right) \delta_j,\end{aligned}$$

we see that

$$\delta'_p \delta'_q = \delta'_q \delta'_p + \sum_k \gamma_k^{(p,q)} \delta'_k \quad (\gamma_k^{(p,q)} \in \mathfrak{F}).$$

Hence $A(y)$ may be expressed as a polynomial, with coefficients in \mathfrak{F} , in the quantities $\delta_1^{i_1} \dots \delta_m^{i_m} y$:

$$A(y) = P(\dots, \delta_1^{i_1} \dots \delta_m^{i_m} y, \dots).$$

Letting the symbols $c_{i_1 \dots i_m}$ denote constants in \mathfrak{F} such that

$$P(\dots, c_{i_1 \dots i_m}, \dots) \neq 0,$$

and letting $\bar{a}_1, \dots, \bar{a}_m$ be unknown constants (that is, indeterminates all of whose derivatives are zero), form the expression

$$\eta = \sum \frac{c_{h_1 \dots h_m}}{h_1! \dots h_m!} (\xi_1 - \bar{a}_1)^{h_1} \dots (\xi_m - \bar{a}_m)^{h_m}.$$

By the above, η satisfies the congruences

$$\delta_1^{i_1} \dots \delta_m^{i_m} \eta \equiv c_{i_1 \dots i_m} \quad (\xi_1 - \bar{a}_1, \dots, \xi_m - \bar{a}_m).$$

Hence

$$A(\eta) \equiv P(\dots, c_{i_1 \dots i_m}, \dots) \quad (\xi_1 - \bar{a}_1, \dots, \xi_m - \bar{a}_m),$$

that is, $A(\eta)$ is a polynomial in the indeterminates $\bar{a}_j = \xi_j - \bar{a}_j$ ($j = 1, \dots, m$) with coefficients in \mathfrak{F} , and these coefficients are not all zero. Therefore we may choose rational values a_j for the unknown constants \bar{a}_j so that, for

$$\eta = \sum \frac{c_{h_1 \dots h_m}}{h_1! \dots h_m!} (\xi_1 - a_1)^{h_1} \dots (\xi_m - a_m)^{h_m},$$

we have $A(\eta) \neq 0$, q.e.d.

4. Existence of a primitive element

We are now in a position to prove our principal

THEOREM. *Let \mathfrak{F} contain m elements whose Jacobian is different from zero. If $\mathfrak{F}(\alpha_1, \dots, \alpha_n)$ is a differential extension field of \mathfrak{F} such that each α_i is a solution of a nonzero differential polynomial in $\mathfrak{F}\{y\}$, then there exists a primitive element γ :*

$$\mathfrak{F}(\alpha_1, \dots, \alpha_n) = \mathfrak{F}(\gamma).$$

By §2 we must show that there exists a $\gamma \in \mathfrak{F}(\alpha_1, \dots, \alpha_n)$ which is invariant

under no isomorphism of $\mathfrak{F}\langle\alpha_1, \dots, \alpha_n\rangle$ with respect to \mathfrak{F} . We shall prove, as a lemma, a stronger result.

Let $A_i(y_i) \in \mathfrak{F}\{y_i\}$ have the solution $y_i = \alpha_i$ ($i = 1, \dots, n$). We shall show that *there exist elements* $\tau_1, \dots, \tau_n \in \mathfrak{F}$ *such that* $\tau_1 y_1 + \dots + \tau_n y_n$ *assumes different values for different solutions of* $\{A_1(y_1), \dots, A_n(y_n)\}$.⁶ Then certainly the element $\tau_1 \alpha_1 + \dots + \tau_n \alpha_n$ will satisfy our requirements on γ .

To prove this lemma, let $z_1, \dots, z_n, t_1, \dots, t_n$ be new unknowns, and, in $\mathfrak{F}\{y_1, \dots, y_n, z_1, \dots, z_n, t_1, \dots, t_n\}$, consider the perfect differential ideal

$$\Omega = \{A_1(y_1), \dots, A_n(y_n), A_1(z_1), \dots, A_n(z_n), t_1(y_1 - z_1) + \dots + t_n(y_n - z_n)\}.$$

Let $\Omega = \Omega_1 \cap \dots \cap \Omega_s$ be the decomposition of Ω into essential prime differential ideals, and suppose the subscripts have been assigned so that $\Omega_1, \dots, \Omega_r$ each contains every $y_i - z_i$, whereas $\Omega_{r+1}, \dots, \Omega_s$ each fails to contain some $y_i - z_i$. Consider an Ω_j with $j > r$. Let $\eta_1, \dots, \eta_n, \bar{\xi}_1, \dots, \bar{\xi}_n, \bar{\tau}_1, \dots, \bar{\tau}_n$ be a generic solution of Ω_j . Since $\bar{\tau}_1(\eta_1 - \bar{\xi}_1) + \dots + \bar{\tau}_n(\eta_n - \bar{\xi}_n) = 0$, and some $\eta_i - \bar{\xi}_i$ is different from zero, $\bar{\tau}_1, \dots, \bar{\tau}_n$ are dependent⁷ over $\mathfrak{F}\langle\eta_1, \dots, \eta_n, \bar{\xi}_1, \dots, \bar{\xi}_n\rangle$. But each η_i and each $\bar{\xi}_i$ annul a nonzero differential polynomial with coefficients in \mathfrak{F} . Hence $\bar{\tau}_1, \dots, \bar{\tau}_n$ are dependent over \mathfrak{F} ,⁸ so that Ω_j contains a nonzero differential polynomial $L_j \in \mathfrak{F}\{t_1, \dots, t_n\}$. Now let $M(t_1, \dots, t_n) = L_{r+1} \dots L_s$. By the authority of §3 choose elements τ_1, \dots, τ_n for which $M(\tau_1, \dots, \tau_n) \neq 0$. For any two distinct solutions $y_i = \eta_i$ ($i = 1; \dots, n$) and $y_i = \zeta_i$ ($i = 1, \dots, n$) of $\{A_1(y_1), \dots, A_n(y_n)\}$, the $3n$ elements

$$\eta_1, \dots, \eta_n, \zeta_1, \dots, \zeta_n, \tau_1, \dots, \tau_n$$

cannot be a solution of Ω . For, these elements cannot be a solution of any Ω_j with $j \leq r$ as each such Ω_j contains every $y_i - z_i$, and they cannot be a solution of an Ω_j with $j > r$ as each such Ω_j contains $M(t_1, \dots, t_n)$. Consequently

$$\tau_1(\eta_1 - \zeta_1) + \dots + \tau_n(\eta_n - \zeta_n) \neq 0.$$

Since η_1, \dots, η_n and ζ_1, \dots, ζ_n were chosen as *any* two distinct solutions of $\{A_1(y_1), \dots, A_n(y_n)\}$, the proof of the lemma, and therefore of the theorem, is complete.

INSTITUTE FOR ADVANCED STUDY

⁶ We lean heavily here on the proof for the ordinary case given by J. F. Ritt, *Differential equations from the algebraic standpoint*, American Mathematical Society Colloquium Publications, vol. XIV, New York, 1932. See especially pp. 26-31.

⁷ See Raudenbush, loc. cit.

⁸ See Raudenbush, loc. cit.

LE CORRESPONDANT TOPOLOGIQUE DE L'UNICITÉ DANS LA THÉORIE DES ÉQUATIONS DIFFÉRENTIELLES¹

PAR N. ARONSZAJN

(Received December 13, 1940; revised July 25, 1942)

Dans la théorie des équations différentielles, aussi bien ordinaires qu'aux dérivées partielles, on a pu établir des théorèmes d'existence et des théorèmes d'unicité. Il est apparu dans beaucoup de cas que, si pour les théorèmes d'existence il suffisait d'admettre pour les membres de l'équation des hypothèses de régularité très faibles, se réduisant parfois à la continuité seule (comme dans le cas de systèmes d'équations différentielles ordinaires), il était nécessaire d'admettre des hypothèses de régularité plus fortes pour assurer l'unicité.

La question se pose de caractériser dans les cas de multiplicité provenant de l'affaiblissement des hypothèses de régularité, l'ensemble des solutions multiples. Il apparaît immédiatement que cette caractérisation doit tenir compte des propriétés topologiques de l'ensemble en question et que pour cela il est nécessaire d'introduire une topologie dans cet ensemble.

Sur cette voie nous sommes arrivé à établir une classe d'ensembles à laquelle appartiennent tous les ensembles des solutions multiples correspondant aux équations en question. Il nous semble probable que tout ensemble de cette classe est homéomorphe à l'ensemble des solutions multiples d'une équation du type considéré. Si cette suggestion était vraie, nous aurions eu ainsi une caractérisation topologique complète de ces ensembles de multiplicité et, en même temps, le correspondant topologique de l'unicité dans le cas de certaines types d'équations admettant de solutions multiples.

Les ensembles de la classe mentionnée seront désignés par R_s . Ce sont des limites des suites décroissantes des ensembles R , ou par R nous désignons les retracts absolus de K. Borsuk.² Les R_s conservent beaucoup de propriétés des retracts absolus.

Notre résultat principal peut être énoncé de manière intuitive (mais peu précise) comme suit: *Si les membres de l'équation en question peuvent être approchés aussi près que l'on veut par les membres d'une équation plus régulière, admettant une solution unique, l'ensemble des solutions de la première équation est un R_s .*

Remarquons que, dans les cas particuliers que nous avons pu traiter, l'hypothèse de notre théorème concernant l'approximation avait pu être vérifiée grâce au théorème de Weierstrass sur l'approximation d'une fonction continue par des

¹ Cet article forme un développement d'une conférence que l'auteur a faite le 19 avril à Paris, à une séance de la Société Math. de France. Les circonstances anormales actuelles n'ont pas permis de donner à cet article un développement aussi complet que l'auteur l'aurait souhaité. Surtout le côté bibliographique est en défaut, mais l'auteur n'a pas pu faire mieux et il s'en excuse.

² Voir au sujet des retracts les articles de K. Borsuk dans *Fundamenta Math.* à partir du t. 17 (1931) pp. 152-170.

polynômes, ou grâce aux théorèmes similaires. A ce propos, relevons que l'application de ce théorème de Weierstrass a déjà été faite par U. Müller³ dans le cas de systèmes d'équations différentielles ordinaires, pour démontrer un théorème de H. Kneser. Ce dernier théorème, qui concerne le caractère continu de l'ensemble de solutions, est une simple conséquence de notre théorème (car R_δ est toujours un continu).

Le travail se compose de quatre paragraphes. Dans le §1 nous rappelons certains résultats et définitions essentiels. Dans le §2 nous prouvons un théorème auxiliaire concernant les suites de rétractes absolus. Le §3 est consacré au résultat fondamental de l'exposé. Des applications aux systèmes d'équations différentielles ordinaires forment le contenu du §4.

1. Résultats Préliminaires

D'après Borsuk² un rétracte absolu (R) est un espace métrique séparable qui est un rétracte de tout espace métrique qui le contient. Les rétractes absolus ont la propriété du point fixe, c'est-à-dire que toute représentation de R sur R possède un point invariant. Nous introduisons la notation R_δ pour désigner tout homéomorphe de l'intersection d'une suite décroissante de rétractes absolus. On peut aisément montrer que l'ensemble R_δ est un continuum à homologie et groupes fondamentaux ceux d'un point. Bien entendu, on sait que ces propriétés appartiennent aussi à R . Cependant R_δ et R peuvent différer en ce qui concerne leurs propriétés locales. Par exemple R_δ peut ne pas avoir de connexions locales, ainsi que le montre clairement l'exemple classique $y = \sin^2(\pi/x)$ pour $0 < x \leq 1$ et $-1 \leq y \leq 1$ pour $x = 0$. Nous observons entre parenthèses qu'un R_δ dans le plan euclidien ne coupe pas le plan.

Dans le but de fournir des conclusions générales, nos résultats sont formulés pour certaines équations opérationnelles de la forme

$$W = T(z)$$

dans les espaces de Banach.⁴ Si T est continu et représente des ensembles bornés de E sur les ensembles (conditionnellement) compacts de E' , on dit alors que T est complètement continu. Notre contribution principale, le théorème C, est basée sur un théorème général d'existence dû à Schauder.⁵

THÉORÈME A. *Lorsque T est complètement continu et représente K sur K , où K est borne, convexe et fermé dans E , son ensemble de points fixes est un sous-ensemble compact en soi et non vide d'un R .*

L'équivalence avec la formulation de Schauder résulte du fait qu'un compactum convexe (ici l'extension convexe fermée⁶ de $T(K)$), dans un espace de

³ Voir M. Müller, *Math. Zeitschrift*, 28 (1928) pp. 619-645.

⁴ S. Banach: *Théorie des Opérations Linéaires*, Warsaw 1932.

⁵ Voir *Math. Zeitschrift*, 26 (1927) pp. 46-65 et *Studia Math.*, 1 et 2. Des théorèmes de ce type ont déjà été donnés par G. D. Birkhoff et O. D. Kellogg, *Transactions Amer. Math. Soc.*, 23 (1925) pp. 96-115; Lefschetz: *Topology* (New York 1930) p. 358 et *Annals of Mathematics*, vol. 38 (1937), pp. 819-822.

⁶ S. Mazur: *Studia*, 2, (1930), pp. 7-10.

Banach est un R . Il serait intéressant de savoir si le théorème C, peut être étendu aux cas où les théorèmes d'existence (fondamentaux) sous-jacents sont démontrés par les méthodes, de Leray-Schauder.⁷

2. Les Suites de Rétractes Absolus

THÉORÈME B. Soit $\{R^{(n)}\}$ une suite de rétractes absolus, sous-ensembles d'un même espace, et soit M un ensemble contenu dans tous les $R^{(n)}$. Si les $R^{(n)}$ convergent vers M , ce dernier ensemble est un R_δ .

Démonstration. Soit \mathfrak{E} l'espace contenant tous les $R^{(n)}$ et soit φ_n une fonction rétractant \mathfrak{E} sur $R^{(n)}$.

Nous pouvons toujours supposer que l'espace \mathfrak{E} est *distanciable* et que l'on a choisi pour lui une distance $\rho(x, y)$ bornée supérieurement (autrement nous aurions pu remplacer \mathfrak{E} par la somme de tous les $R^{(n)}$ qui a certainement ces propriétés).

Nous allons choisir une sous-suite $\{\bar{R}^{(k)}\}$ de $\{R^{(n)}\}$ de sorte que, en désignant par $\bar{\varphi}_k$ la fonction φ_n correspondant à $\bar{R}^{(k)}$, et par $\psi_i^{(k)}$, pour $i < k$, la fonction composée

$$(1) \quad \psi_i^{(k)} = \bar{\varphi}_i \bar{\varphi}_{i+1} \cdots \bar{\varphi}_{k-1},$$

on ait pour tous $k, i < k$ et $x \in \bar{R}^{(k)}$,

$$(2) \quad \rho(x, \psi_i^{(k)}(x)) \leq 1/k.$$

Pour définir les $\bar{R}^{(k)}$ nous commençons par poser $\bar{R}^{(1)} = R^{(1)}$. Supposons maintenant que les $\bar{R}^{(1)}, \bar{R}^{(2)}, \dots, \bar{R}^{(k)}$ sont déjà définis. D'après (1), $\psi_i^{(k+1)}$ est alors défini, et on a pour tout x de M et tout $i < k + 1$

$$x = \psi_i^{(k+1)}(x),$$

car pour tout r , $M \subset \bar{R}^{(r)}$, et par conséquent $\psi_r(x) = x$. Il s'ensuit qu'il existe un voisinage V de M tel que, pour $x \in V$ et tout $i < k + 1$, on ait,

$$\varphi(x, \psi_i^{(k+1)}(x)) \leq \frac{1}{k+1}.$$

Nous poserons $\bar{R}^{(k+1)} =$ le premier $R^{(n)}$ postérieur à $\bar{R}^{(k)}$ dans la suite $\{R^{(n)}\}$, contenu dans V . Il est clair qu'un tel $R^{(n)}$ existe vu que les $R^{(n)}$ convergent vers M . Ainsi, les $\bar{R}^{(k)}$ se définissent successivement et la propriété (2) est remplie.

Considérons maintenant le produit combinatoire infini $\mathfrak{E}^\infty = \mathfrak{E} \times \mathfrak{E} \times \dots$ aux éléments $X = (x_1, x_2, \dots, x_n, \dots)$ avec $x_n \in \mathfrak{E}$. On définit dans \mathfrak{E}^∞ une distance à la Fréchet

$$\rho(X, Y) = \sum_{n=1}^{\infty} 2^{-n} \rho(x_n, y_n).$$

⁷ Voir J. Leray et J. Schauder, Annales Scient. École Norm. Sup., 51 (1934) pp. 45-78.

La notion de limite correspondante se définit comme suit: la suite $\{X^{(k)}\}$ converge vers X , si chaque suite $\{x_n^{(k)}\}$ converge vers x_n .

Considérons dans \mathcal{E}^∞ les sous-ensembles $Q^{(k)}$, définis de manière suivante: $Q^{(k)}$ est composé de tous les points $X = (x_1, x_2, \dots, x_k, \dots)$ tels que, pour $n \geq k$, $x_n \in \bar{R}^{(n)}$, tandis que pour $n < k$, $x_n = \psi_n^{(k)}(x_k)$.

Il est clair que $Q^{(k)}$ est homéomorphe avec l'ensemble de toutes les suites (x_k, x_{k+1}, \dots) où x_n parcourt $\bar{R}^{(n)}$, $n = k, k+1, \dots$. Cet ensemble forme le produit combinatoire $\bar{R}^{(k)} \times \bar{R}^{(k+1)} \times \dots$ de rétractes absolus $\bar{R}^{(n)}$; c'est donc un rétracte absolu.⁸ Il en résulte que $Q^{(k)}$ est un rétracte absolu.

Remarquons ensuite que $Q^{(k)} \supset Q^{(k+1)}$. En effet, si $X = (x_1, x_2, \dots, x_k, x_{k+1}, \dots)$ appartient à $Q^{(k+1)}$, on a d'après la définition de $Q^{(k+1)}$: $x_n \in \bar{R}^{(n)}$ pour $n \geq k+1$, $x_k = \psi_k^{(k+1)}(x_{k+1}) = \bar{\varphi}_k(x_{k+1}) \in \bar{R}^{(k)}$ et enfin, pour $n < k$, $x_n = \psi_n^{(k+1)}(x_{k+1}) = \bar{\varphi}_n \bar{\varphi}_{n+1} \dots \bar{\varphi}_k(x_{k+1}) = \psi_n^{(k)} \bar{\varphi}_k(x_{k+1}) = \psi_n^{(k)}(x_k)$, donc $X \in Q^{(k)}$.

Prouvons maintenant que la suite décroissante $Q^{(1)}, Q^{(2)}, \dots$ a pour intersection l'ensemble M' composé de tous les $X = (x_1, x_2, \dots)$ avec $x_1 = x_2 = x_3 = \dots = x \in M$. En effet, si $X = (x_1, x_2, \dots)$ appartient à tous les $Q^{(k)}$, on aura suivant (2) pour tout k et tout $n < k$

$$\rho(x_k, x_n) = \rho(x_k, \psi_n^{(k)}(x_k)) \leq \frac{1}{k}.$$

Il en résulte que tout x_n , $n = 1, 2, \dots$, est la limite de la suite $\{x_k\}$ qui est nécessairement convergente. Il s'ensuit d'une part que $x_1 = x_2 = \dots = x$. D'autre part, $x_k \in \bar{R}^{(k)}$ et, les $\bar{R}^{(k)}$ convergeant vers M , la limite x de $\{x_k\}$ appartient à M . Ainsi $M' \supset Q^{(1)}Q^{(2)} \dots$. Inversement, si $X \in M'$, il appartient à tout $Q^{(k)}$, car, pour $n \geq k$, $x_n = x \in M \subset \bar{R}^{(n)}$ et, pour $n < k$, $x_n = x = \psi_n^{(k)}(x) = \psi_n^{(k)}(x_k)$, vu que toute φ_i transforme un $x \in M$ en lui-même. Il est donc prouvé que $M' = Q^{(1)}Q^{(2)} \dots$.

Enfin, il est évident que l'ensemble M' est homéomorphe avec M par l'intermédiaire de la correspondance donnant à un $X = (x, x, x, \dots)$ de M' , pour image le point x de M .

Ainsi, M est homéomorphe avec M' qui est, d'après ce qui précède, un R_δ . M est donc lui-même aussi un R_δ , c.q.f.d.

3. Le théorème principal

Pour pouvoir poursuivre nos raisonnements nous allons admettre que la transformation T peut être approchée aussi près que l'on veut par une transformation "plus régulière."

Pour préciser cette hypothèse revenons aux notations du §1 précédent. Nous supposons qu'à tout $\epsilon > 0$ on peut faire correspondre une transformation T_ϵ complètement continue de l'espace E en lui-même de sorte que

1°. $\|T_\epsilon(z) - T(z)\| \leq \epsilon$ pour tout élément z de K , $\|\cdot\|$ désignant la norme dans E ;

⁸ N. Aronszajn et K. Borsuk, Fundamenta, 18, 1932, pp. 193-197.

2°. La transformation $z_1 = z - T_\epsilon(z) \equiv H_\epsilon(z)$ représente de manière biunivoque l'ensemble K en un ensemble contenant une sphère $\|z\| \leq \rho$, avec ρ indépendant de ϵ .

Tandis que la première condition précise de manière dont T est approchée par T_ϵ , la seconde condition peut être appliquée "condition d'unicité," car elle a pour conséquence l'existence et l'unicité (si l'on se limite aux solutions appartenant à K) de la solution de l'équation en z

$$z - T_\epsilon(z) = z_1,$$

pour z_1 de norme suffisamment petite. Comme nous l'avons déjà remarqué, pour avoir l'unicité de solution il faut admettre en général des conditions supplémentaires de régularité, c'est pourquoi nous dirons que T_ϵ est "plus régulière" que T .

Dans ces conditions nous pouvons démontrer le

THÉORÈME C. L'ensemble des solutions de l'équation $T(z) = z$ est un R_δ .

DÉMONSTRATION. Désignons cet ensemble des solutions par S . Considérons les transformations $T_n \equiv T_{\epsilon_n}$ pour une suite $\{\epsilon_n\}$ tendant vers 0, tous les ϵ_n étant $\leq \rho$.

Considérons d'abord les transformations T_n et H_n pour un n fixe. L'ensemble S étant contenu dans K , on a d'après 1°, pour tout élément ζ de S ,

$$\|H_n(\zeta)\| = \|H_{\epsilon_n}(\zeta)\| = \|\zeta - T_n(\zeta)\| = \|T(\zeta) - T_n(\zeta)\| \leq \epsilon_n.$$

Par conséquent, l'ensemble transformé $H_n(S)$ est contenu dans la sphère de rayon $\epsilon_n \leq \rho$. Cet ensemble, obtenu par transformation continue d'un ensemble compact en soi, est compact en soi (voir le théorème A). Le plus petit corps convexe le contenant est aussi compact en soi et compris dans la sphère de rayon $\epsilon_n \leq \rho$.

Soit Q_n ce corps convexe. D'après la condition 2°, la transformation H_n^{-1} inverse de la transformation H_n , est définie sur Q_n . Elle donne de Q_n une image $R^{(n)} = H_n^{-1}(Q_n)$ contenue dans l'ensemble K .

La transformation inverse H_n^{-1} n'est pas en général continue, mais nous allons montrer qu'elle l'est sur Q_n . En effet, si une suite d'éléments $\{h_k\}$ de Q_n tend vers h (qui appartient aussi à Q_n , celui-ci étant compact en soi), on a pour les éléments $z_k = H_n^{-1}(h_k)$ les équations suivantes

$$z_k - T_n(z_k) = h_k.$$

Les z_k appartenant à l'ensemble borné K , ils forment une suite bornée, et la transformation complètement continue T_n transforme cette suite en une suite compacte. Si les z_k ne tendaient pas vers l'élément $z = H_n^{-1}(h)$ donné par l'équation

$$z - T_n(z) = h,$$

on pourrait extraire des z_k une suite $\{z_{k_i}\}$ n'admettant pas z comme élément limite et telle que les $T_n(z_{k_i})$ convergent vers un élément g . Mais alors les

éléments $z_{k_i} = T_n(z_{k_i}) + h_{k_i}$ convergeraient vers $g + h$ et les $T_n(z_{k_i})$ convergeraient vers $T_n(g + h)$, qui serait égal à g , et on aurait

$$g + h = T_n(g + h) + h;$$

$g + h$ serait donc la solution z de $z - T_n(z) = h$ et les z_{k_i} y convergeraient, d'où contradiction.

Cette contradiction prouve que, sur Q_n , la transformation H_n^{-1} est continue. Puisque son inverse H_n est d'après 2° biunivoque et continue, la transformation H_n^{-1} représente de manière *homéomorphe* Q_n en $R^{(n)} = H_n^{-1}(Q_n)$.

L'ensemble Q_n étant compact en soi et convexe, c'est un *rétracte absolu*. Par conséquent, son homéomorphe $R^{(n)}$ l'est aussi. D'autre part Q_n contenait le transformé $H_n(S)$ de S . Il s'ensuit que $R^{(n)} = H_n^{-1}(Q_n)$ contient S . Dès lors, pour prouver notre théorème, il nous reste à prouver que les $R^{(n)}$ convergent vers S pour n tendant vers l' ∞ .

A cet effet prenons une suite quelconque $\{z_j\}$ telle que chaque z_j appartient à un $R^{(n_j)}$, les n_j tendant vers l' ∞ . Comme nous l'avons vu plus haut, tous les $R^{(n)}$ sont contenus dans l'ensemble borné K . Il s'ensuit que $\{z_j\}$ est une suite bornée et que les $T(z_j)$ forment une suite compacte de laquelle on peut extraire une sous-suite $\{T(z_{j_k})\}$ convergeant vers un élément z . Les z_j appartenant à K , on a selon 1°

$$\|T_{n_{j_k}}(z_{j_k}) - T(z_{j_k})\| \leq \epsilon_{n_{j_k}} \rightarrow 0,$$

donc, les $T_{n_{j_k}}(z_{j_k})$ convergent aussi vers z . D'après la définition des $R^{(n)}$, on a pour tout z_{j_k} l'équation

$$z_{j_k} = T_{n_{j_k}}(z_j) + h_{j_k},$$

où h_{j_k} appartient à $Q_{n_{j_k}}$, donc à une sphère de rayon $\epsilon_{n_{j_k}} \rightarrow 0$. Il en résulte successivement: $\lim z_{j_k} = \lim T_{n_{j_k}}(z_{j_k}) = z$, $\lim T(z_{j_k}) = T(z)$ donc $T(z) = z$.

Ainsi; de toute suite $\{z_j\}$ avec z_j appartenant à $R^{(n_j)}$ on peut extraire une suite $\{z_{j_k}\}$ convergeant vers une solution z de l'équation $T(z) = z$, donc vers un élément de S . Ceci prouve que les $R^{(n)}$ convergent vers S .

Notre théorème est ainsi démontré.

4. Application

Comme application de notre théorème général nous allons considérer un système d'équations différentielles ordinaires. Sans restreindre essentiellement la généralité nous pouvons nous limiter au cas d'un système de deux équations avec deux fonctions inconnues

$$(3) \quad \frac{dx}{dt} = u(x, y, t), \quad \frac{dy}{dt} = v(x, y, t).$$

avec les conditions initiales

$$x(0) = 0, \quad y(0) = 0.$$

Il est connu depuis Peano que ce système admet certainement de solutions, si seulement u et v sont continues. En général ces solutions existeront dans un intervalle de t entourant $t = 0$. Si nous supposons—ce que nous allons faire dans la suite—que les fonctions u et v sont continues et bornées pour toutes les valeurs de x , y et t , les solutions existeront sur tout l'axe de t .

D'autre part, on sait que si l'on suppose les fonctions u et v satisfaisant à une condition de Lipschitz relativement à x et y , uniformément en t dans tout intervalle fini, la solution est unique. Elle le sera donc à fortiori, si u et v sont analytiques en x , y et t , ou ne diffèrent d'une telle fonction que par une fonction de t seul.

Pour appliquer notre théorème général, nous envisagerons l'espace vectoriel de tous les couples de fonctions $[x(t), y(t)]$ admettant des dérivées $x'(t)$ et $y'(t)$ continues, et satisfaisant aux conditions $x(0) = y(0) = 0$. Nous considérerons ces fonctions dans un intervalle fini fixe $\alpha \leq t \leq \beta$, $\alpha < 0 < \beta$, arbitrairement choisi.

Dans cet espace vectoriel nous prendrons comme norme d'un couple de fonctions

$$z = [x(t), y(t)],$$

le nombre

$$\|z\| = \max_{\alpha \leq t \leq \beta} [x'(t)^2 + y'(t)^2]^{\frac{1}{2}}.$$

Considérons dans cet espace la transformation

$$z_1 = T(z) = [x_1(t), y_1(t)], \quad x_1(t) = \int_0^t u(x, y, t) dt, \quad y_1(t) = \int_0^t v dt,$$

où $z = [x(t), y(t)]$.

Il est clair que, si l'on pose

$$m = \text{borne sup } (u^2 + v^2)^{\frac{1}{2}}, \text{ pour tous les } x, y, t,$$

la sphère de notre espace vectoriel,

$$\|z\| \leq 2m$$

peut être prise comme l'ensemble K dans la théorie générale, car la transformation T la représente en elle-même. D'autre part, on prouve facilement que T est complètement continue. Ceci permet déjà d'appliquer le théorème d'existence A.

Pour appliquer le théorème C, il faut définir les transformations T , conformément aux conditions 1° et 2° du §3. A cet effet remarquons d'abord que, si

pour $z = [x(t), y(t)]$ on a $\|z\| \leq 2m$, il en résulte pour $x(t)$ et $y(t)$, $\alpha \leq t \leq \beta$, les inégalités

$$|x(t)| \leq 2m(\beta - \alpha), \quad |y(t)| \leq 2m(\beta - \alpha).$$

Par conséquent, pour satisfaire aux conditions 1° et 2°, il suffira d'approcher chacune des fonctions $u(x, y, t)$ et $v(x, y, t)$ par des fonctions analytiques u_ϵ et v_ϵ des trois variables réelles x, y et t , satisfaisant aux inégalités

$$|u_\epsilon - u| \leq \frac{\epsilon}{2}, \quad |v_\epsilon - v| \leq \frac{\epsilon}{2}, \text{ pour}$$

$$|x| \leq 2m(\beta - \alpha), \quad |y| \leq 2m(\beta - \alpha), \quad \alpha \leq t \leq \beta,$$

$$|u_\epsilon| \leq m\sqrt{2}, \quad |v_\epsilon| \leq m\sqrt{2} \text{ pour tous les } x, y, t.$$

En se basant sur le théorème d'approximation de Weierstrass on construit aisément les fonctions u_ϵ et v_ϵ . Les théorèmes d'existence et d'unicité, indiqués au commencement de ce § permettent de vérifier immédiatement la condition 2°. Ainsi, le théorème C est applicable.

Indiquons quelques conséquences de ce théorème dans le cas présent. Comme on sait, à chaque solution de notre système avec les conditions initiales $x(0) = y(0) = 0$, correspond dans l'espace des variables x, y, t une courbe intégrale du système passant par l'origine $x = y = t = 0$. S'il y a plusieurs solutions, il passe par l'origine tout un faisceau de courbes intégrales. Chacune de ces courbes coupe le plan $t = t_0$ au point $x_0 = x(t_0)$, $y_0 = y(t_0)$ qui varie de façon continue quand la courbe intégrale parcourt le faisceau. Il s'ensuit que la trace du faisceau sur le plan $t = t_0$ est une image continue du faisceau, c'est à dire de l'ensemble S des solutions du système (3). Cet ensemble étant un R_1 , donc à fortiori un continu, son image est également un continu. Par conséquent, la trace sur le plan $t = t_0$ du faisceau des courbes intégrales passant par l'origine —ou par un point quelconque—est un continu. C'est le théorème de Kneser; il se montre ainsi une conséquence immédiate de notre théorème.

Dans des cas particuliers nous pouvons préciser la nature de cette trace. Par exemple, si les fonctions $u(x, y, t)$ et $v(x, y, t)$ satisfont dans tout l'espace des x, y, t la même condition de Lipschitz sauf au point $x = y = t = 0$, il n'y aura dans tout l'espace que l'origine comme point par lequel puissent passer plusieurs courbes intégrales. Dans ce cas, chaque point de la trace sur le plan $t = t_0 \neq 0$ ne provient que d'une seule courbe intégrale. Par conséquent, la trace est une image homéomorphe de l'ensemble S et est un R_1 . D'après la propriété caractéristique des R_1 plans, cette trace ne coupe pas le plan $t = t_0$.

Il est très probable que tout R_1 plan peut être obtenu comme trace du faisceau intégral pour un choix convenable des fonctions u et v conformes aux conditions ci-dessus. Ceci est en rapport avec l'hypothèse que nous avons émise dans l'introduction.

Il serait intéressant d'étudier la nature du faisceau intégral et de ses traces pour différentes classes de fonctions. En particulier, on pourrait étudier la nature topologique du faisceau intégral pour les fonctions u et v continues et bornées dans tout l'espace (x, y, t) et analytiques partout sauf à l'origine.

EDITORS NOTE. Owing to present circumstances it was impossible to communicate freely with the author regarding certain necessary revisions in the paper. With the authorization of the author and some information conveyed by him, this was accomplished by Professor D. G. Bourgin, to whom the Editors wish to express their personal thanks and also those of the author. S. I.,

LONDON

A PROOF THAT THERE EXISTS A CIRCUMSCRIBING CUBE AROUND ANY BOUNDED CLOSED CONVEX SET IN R^3

BY SHIZUO KAKUTANI

(Received June 8, 1942)

1

The following problem was proposed by Professor Rademacher: Given a bounded closed convex set in a three-space R^3 , is it always possible to find a circumscribing cube around it? It is easy to see (cf. §3) that this problem can be reduced to the following one: Given a real-valued continuous function $f(P)$ defined on a two-sphere S^2 , is it possible to find a triple of points $P_1, P_2, P_3 \in S^2$, perpendicular to one another (this means that the three vectors $\mathbf{OP}_1, \mathbf{OP}_2, \mathbf{OP}_3$ from the center O of S^2 to these three points P_1, P_2, P_3 are perpendicular to one another) such that $f(P_1) = f(P_2) = f(P_3)$? The purpose of the present note is to answer these questions in the affirmative.

2

THEOREM 1. *Let $f(P)$ be a real-valued continuous function defined on a two-sphere S^2 . Then there exists a triple of points $P_1, P_2, P_3 \in S^2$, perpendicular to one another, such that $f(P_1) = f(P_2) = f(P_3)$.*

PROOF. Let us consider S^2 as a sphere of radius 1 in a three-space R^3 , with the origin $O = (0, 0, 0)$ of R^3 as a center. Let us put $P_1^0 = (1, 0, 0)$, $P_2^0 = (0, 1, 0)$, $P_3^0 = (0, 0, 1)$. Let further $G = \{\sigma\}$ be the group of all rotations of S^2 (or equivalently, rotations of R^3 around its origin $O = (0, 0, 0)$). G is a three dimensional compact manifold.

For any $\sigma \in G$, consider the point $\varphi(\sigma) = (x, y, z) \in R^3$ defined by $x = f(\sigma^{-1}(P_1^0))$, $y = f(\sigma^{-1}(P_2^0))$, $z = f(\sigma^{-1}(P_3^0))$. It is clear that $\sigma \rightarrow \varphi(\sigma)$ is a continuous mapping of G into R^3 . In order to prove our theorem, it suffices to show that there exists a rotation $\sigma \in G$ such that $\varphi(\sigma)$ lies on the straight line $l: x = y = z$ in R^3 .

We assume the contrary, and shall draw a contradiction from it. Let ρ be the projection of R^3 onto the plane $\pi: x + y + z = 0$, which is perpendicular to the line l . Then $\sigma \rightarrow \psi(\sigma) \equiv \rho(\varphi(\sigma))$ is a continuous mapping of G into π . By assumption, the image $\psi(G)$ of G by this mapping $\psi(\sigma)$ does not contain the origin $O = (0, 0, 0)$.

Let H be the subgroup of G consisting of all rotations around the line l . H is isomorphic to the group of rotations of the plane π around the origin $O = (0, 0, 0)$, and we may denote elements of H by $\sigma_\theta (0 \leq \theta \leq 2\pi)$, where θ denotes the angle of rotation around the axis l , measured in such a sense that we have $\sigma_{2\pi/3}(P_i^0) = P_{i+1}^0, i = 1, 2, 3, \text{ mod } 3$.

Let us denote the rotation of the plane π around its origin $O = (0, 0, 0)$, which corresponds to σ_θ , by $\tau_\theta (0 \leq \theta \leq 2\pi)$. It is then easy to see that we have

$$\psi(\sigma_{\theta+2(\pi/3)}) = \tau_{2\pi/3}(\psi(\sigma_\theta)), \quad \psi(\sigma_{\theta+4(\pi/3)}) = \tau_{4\pi/3}(\psi(\sigma_\theta))$$

for any $\theta (0 \leq \theta \leq 2\pi)$. Let C_{θ_1, θ_2} be the curve traced on π by $\psi(\sigma_\theta)$ when θ runs over the interval $\theta_1 \leq \theta \leq \theta_2$. Then the fact stated above means that the curves $C_{2\pi/3, 4\pi/3}$ and $C_{4\pi/3, 2\pi}$ are obtained by applying the rotations $\tau_{2\pi/3}$ and $\tau_{4\pi/3}$ to the curve $C_{0, 2\pi/3}$.

Let α be the increment of the angle around the origin $O = (0, 0, 0)$ in the plane π , when the point $\psi(\sigma_\theta)$ runs over the curve $C_{0, 2\pi/3}$ from $\psi(\sigma_0)$ to $(\sigma_{2\pi/3})$, or equivalently, when θ runs from 0 to $2\pi/3$. Then α must be of the form: $\alpha = 2m\pi + 2\pi/3$, where m is an integer ($m = 0, \pm 1, \pm 2, \dots$). Hence as θ runs from 0 to 2π , the total increment of the angle of $\psi(\sigma_\theta)$ around the origin $O = (0, 0, 0)$ in the plane π is $6m\pi + 2\pi = (3m + 1) \cdot 2\pi$.

On the other hand, consider H as a closed curve on the manifold of the topological group G . Then it is well known that $2H$ is homotopic to zero on G . Consequently, the curve $2C_{0, 2\pi}$, which is the image of $2H$ by the mapping $\sigma \rightarrow \psi(\sigma)$, must also be homotopic to zero on π^* , where by π^* we mean the open set which is obtained by taking away the origin $O = (0, 0, 0)$ from the plane π . This is, however, impossible, since the total increment of the angle on the curve $2C_{0, 2\pi}$ is $2(3m + 1) \cdot 2\pi \neq 0$.

Thus we arrive at a contradiction, and the proof of Theorem 1 is completed.

3

THEOREM 2. *Let K be a bounded closed convex set in a three space R^3 . Then there exists a circumscribing cube around K .*

PROOF. Let S^2 be a two sphere in R^3 with the origin $O = (0, 0, 0)$ of R^3 as a center. For any point $P \in S^2$, consider two tangent planes to K (parallel to each other) which are perpendicular to the vector \overrightarrow{OP} . These two planes may coincide if K is a flat convex set. Let $f(P)$ be the vertical distance of these two planes. $f(P)$ is clearly a real-valued continuous function defined on S^2 . (Moreover, $f(P)$ takes the same value at two antipodal points of S^2 ; but we do not need this fact in our proof). By Theorem 1, there exists a triple of points $P_1, P_2, P_3 \in S^2$, perpendicular to one another, such that $f(P_1) = f(P_2) = f(P_3)$. It is then clear that the corresponding six tangent planes form a cube which is circumscribing around the convex set K .

4. Remarks

There are two problems related to our results. The first one is to investigate whether it is possible to inscribe a cube in a given bounded open convex set in R^3 . The answer to this question is negative, and a counter-example to this is given by a tetrahedron in R^3 which is extremely flat. In fact, if we take a convex quadrangle $ABCD$ on the (x, y) -plane, such that the two diagonals AC and BD are not perpendicular to each other, and if we shift the vertex A in a direction of z -axis by a small distance, then the tetrahedron $A'BCD$ thus obtained is a required one. It is easy to see that there is no inscribing cube in this tetrahedron.

The second problem concerns the possibility of generalizations to higher dimensional cases. It is not yet known whether or not it is possible to find a circumscribing n -dimensional cube around any given bounded closed convex set in R^n ($n \geq 4$). We may also ask: Given a real-valued continuous function $f(P)$ defined on an $(n - 1)$ -sphere S^{n-1} , is it possible to find n points P_1, \dots, P_n on S^{n-1} , perpendicular to one another, such that $f(P_1) = \dots = f(P_n)$ ($n \geq 4$)? These problems are still unsolved.

INSTITUTE FOR ADVANCED STUDY

AN EXTREMUM PROBLEM IN PRODUCT MEASURE

By SHIZUO KAKUTANI

(Received June 19, 1942)

I. The problem and results

The following problem was proposed by P. R. Halmos: Let Φ be the collection of all real valued measurable functions $\varphi(x, y)$ defined on the unit square $I_x \times I_y$: $0 \leq x, y \leq 1$ such that $0 \leq \varphi(x, y) \leq 1$ for any $(x, y) \in I_x \times I_y$. Let us put

$$(1) \quad A(\varphi) = \int_0^1 \int_0^1 \varphi(x, y) dx dy,$$

$$(2) \quad V(\varphi) = \int_0^1 \int_0^1 \int_0^1 \varphi(x, y) \varphi(y, z) dx dy dz,$$

and

$$(3) \quad \lambda(\alpha) = \sup_{\substack{\varphi \in \Phi \\ A(\varphi) = \alpha}} V(\varphi),$$

$$(4) \quad \mu(\alpha) = \inf_{\substack{\varphi \in \Phi \\ A(\varphi) = \alpha}} V(\varphi),$$

where α is a real number ($0 \leq \alpha \leq 1$). Then what are the exact values of $\lambda(\alpha)$ and $\mu(\alpha)$ as functions of α in the interval $0 \leq \alpha \leq 1$?

Consider the special case when $\varphi(x, y)$ is the characteristic function $\varphi_E(x, y)$ of a measurable set $E \subseteq I_x \times I_y$. Then $A(\varphi_E) = A(\varphi)$ is clearly the area (= two dimensional Lebesgue measure) of the set E , while the meaning of $V(\varphi_E) = V(\varphi)$ may be interpreted as follows: Take the unit cube $I_x \times I_y \times I_z$: $0 \leq x, y, z \leq 1$, and consider E as a subset of its face $I_x \times I_y \times (0)$. Let E' be the set on the face $(0) \times I_y \times I_z$ which is obtained from E by the mapping $(x, y, 0) \rightarrow (0, x, y)$. Then $V(\varphi_E)$ is the volume (= three dimensional Lebesgue measure) of the intersection of two cylindrical sets $E \times I_z$ and $I_x \times E'$, i.e., the set of all points $(x, y, z) \in I_x \times I_y \times I_z$ such that $(x, y) \in E$ and $(y, z) \in E'$.

The purpose of the present note is to prove the following

THEOREM.

$$(5) \quad \lambda(\alpha) = 2\alpha - 1 + (1 - \alpha)^{\frac{1}{2}}, \quad 0 \leq \alpha \leq \frac{1}{2}$$

$$(6) \quad \lambda(\alpha) = \alpha^{\frac{1}{2}}, \quad \frac{1}{2} \leq \alpha \leq 1$$

$$(7) \quad \mu(\alpha) = \frac{n-2}{3n^2} \left\{ (3\alpha-1)n + 1 - \frac{((1-2\alpha)n-1)^{\frac{1}{2}}}{(n-1)^{\frac{1}{2}}} \right\},$$

$$\frac{1}{2} \left(1 - \frac{1}{n-1} \right) \leq \alpha \leq \frac{1}{2} \left(1 + \frac{1}{n} \right), \quad n = 2, 3, \dots$$

$$(8) \quad \mu\left(\frac{1}{2}\right) = \frac{1}{3},$$

$$\mu(\alpha) = 2\alpha - 1 + \frac{n-2}{3n^2} \left\{ (2-3\alpha)n + 1 - \frac{((2\alpha-1)n-1)^{\frac{1}{2}}}{(n-1)^{\frac{1}{2}}} \right\},$$

$$(9) \quad \frac{1}{2} \left(1 + \frac{1}{n} \right) \leq \alpha \leq \frac{1}{2} \left(1 + \frac{1}{n-1} \right), \quad n = 2, 3, \dots$$

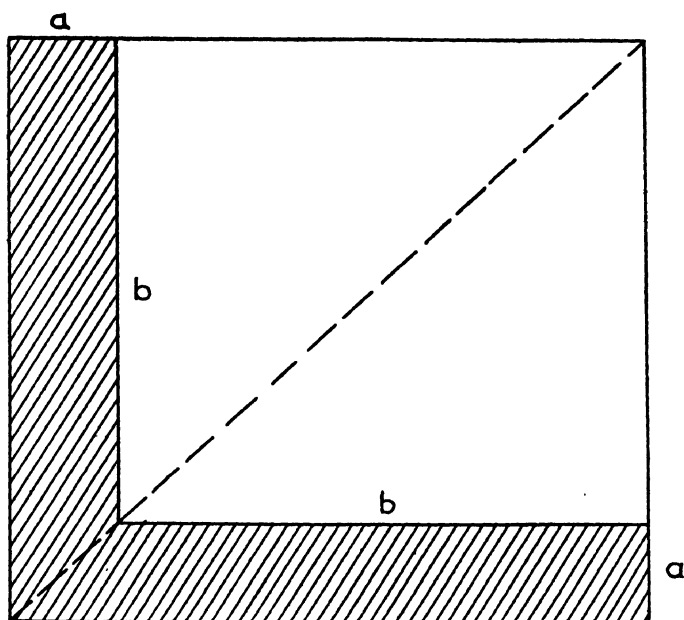


FIG. 1

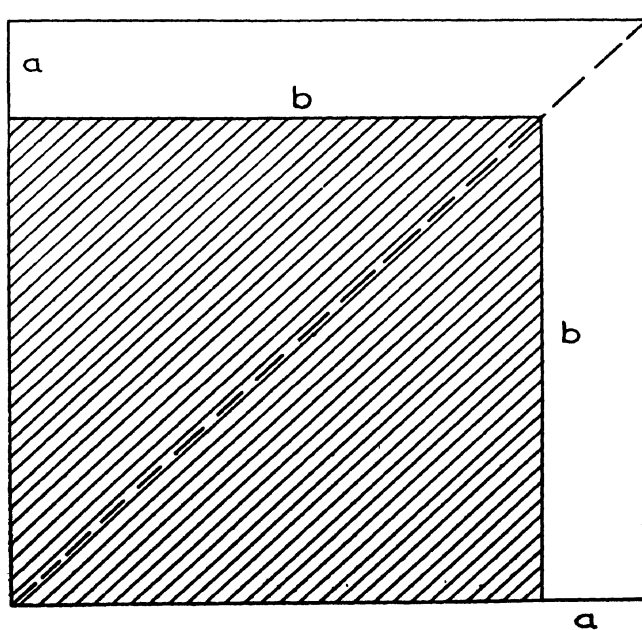


FIG. 2

These extreme values of $V(\varphi)$ are attained by the characteristic functions $\varphi_E(x, y)$ of the sets E of Figs. 1, 2, 3, 4, and 5 respectively, where

$$\bar{a} = \bar{a}' = 1 - (1 - \alpha)^{\frac{1}{2}}, \quad \bar{b} = \bar{b}' = (1 - \alpha)^{\frac{1}{2}} \quad (\text{in Fig. 1});$$

$$\bar{a} = \bar{a}' = 1 - \alpha^{\frac{1}{2}}, \quad \bar{b} = \bar{b}' = \alpha^{\frac{1}{2}} \quad (\text{in Fig. 2});$$

$$\bar{a}_1 = \bar{a}'_1 = \cdots = \bar{a}_{n-1} = \bar{a}'_{n-1} = \frac{1}{n} \left(1 + \left(\frac{n\beta - 1}{n-1} \right)^{\frac{1}{2}} \right)$$

$$\bar{a}_n = \bar{a}'_n = \frac{1}{n} (1 - ((n-1)(n\beta - 1))^{\frac{1}{2}})$$

where n is a positive integer satisfying

$$\frac{1}{n} \leq \beta \equiv |1 - 2\alpha| \leq \frac{1}{n-1} \quad (\text{in Figs. 3, 5}).$$

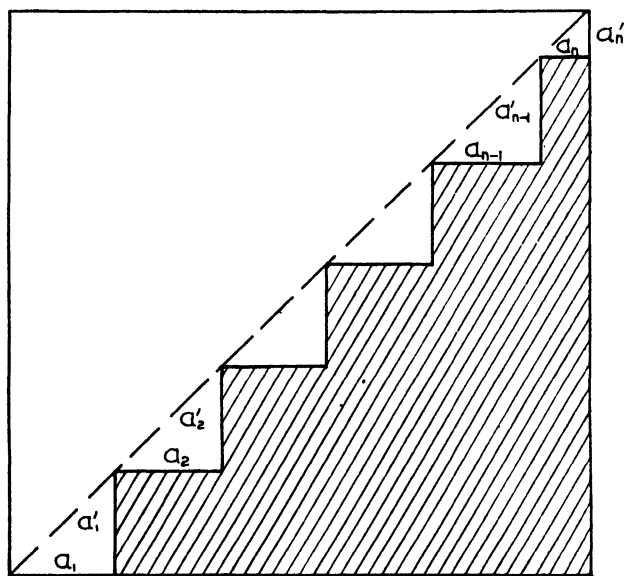


FIG. 3

The graphs of $\lambda(\alpha)$ and $\mu(\alpha)$ are given in Figs. 6 and 7. It is to be remarked that $\lambda''(\alpha) > 0$ in the intervals $0 < \alpha < \frac{1}{2}$ and $\frac{1}{2} < \alpha < 1$, while at $\alpha = \frac{1}{2}$ we have $\lambda'(\frac{1}{2} - 0) = 0.94 \cdots < \lambda'(\frac{1}{2} + 0) = 1.05 \cdots$. $\mu(\alpha)$ is linear in the intervals $0 \leq \alpha \leq \frac{1}{2}$ and $\frac{1}{2} \leq \alpha \leq 1$. Further we have $\mu''(\alpha) < 0$ in the intervals $\frac{1}{2} \left(1 - \frac{1}{n-1} \right) < \alpha < \frac{1}{2} \left(1 + \frac{1}{n} \right)$ and $\frac{1}{2} \left(1 + \frac{1}{n} \right) < \alpha < \frac{1}{2} \left(1 - \frac{1}{n-1} \right)$, $n = 3, 4, \dots$. Finally, at $\alpha = \frac{1}{2} \left(1 \pm \frac{1}{n} \right)$ we have

$$\mu' \left(\frac{1}{2} \left(1 - \frac{1}{n} \right) - 0 \right) = \frac{n-2}{n} < \mu' \left(\frac{1}{2} \left(1 - \frac{1}{n} \right) + 0 \right) = \frac{n-1}{n},$$

$$\mu' \left(\frac{1}{2} \left(1 + \frac{1}{n} \right) - 0 \right) = \frac{n+1}{n} < \mu' \left(\frac{1}{2} \left(1 + \frac{1}{n} \right) + 0 \right) = \frac{n+2}{n}.$$

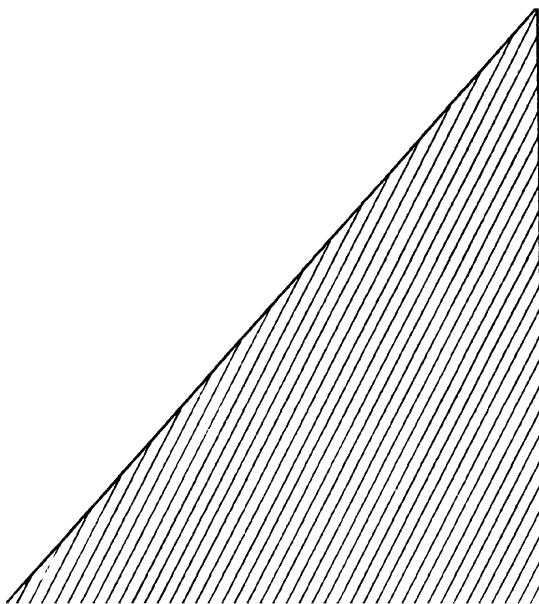


FIG. 4

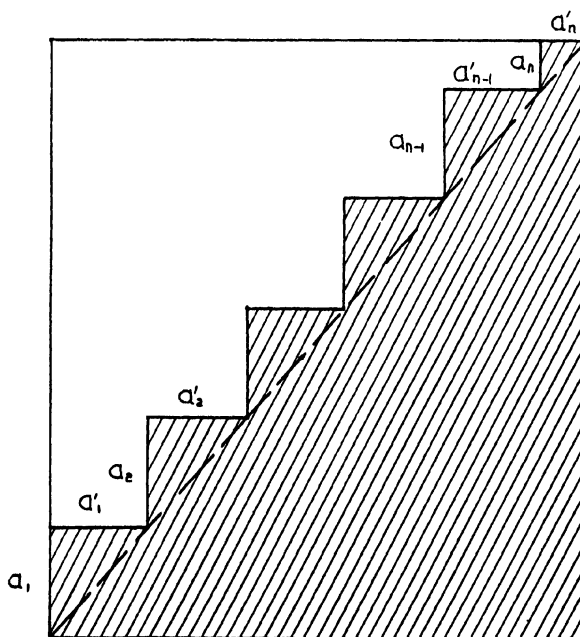


FIG. 5

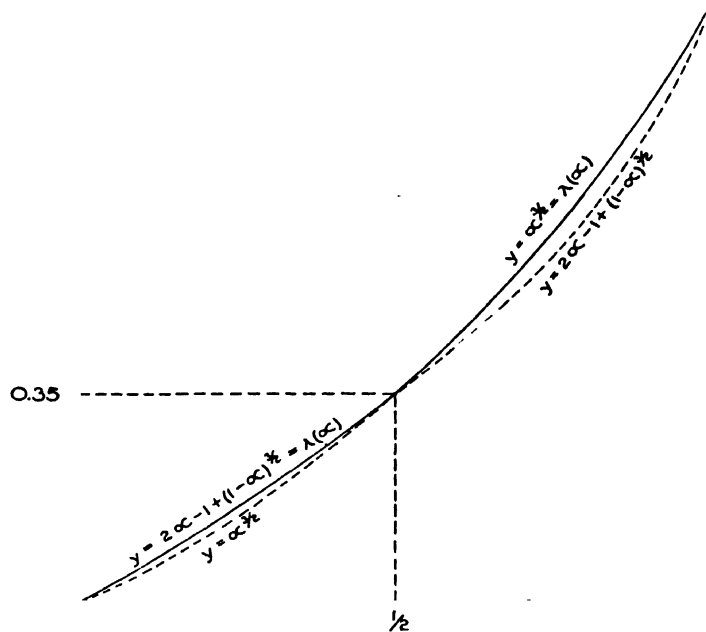


FIG. 6

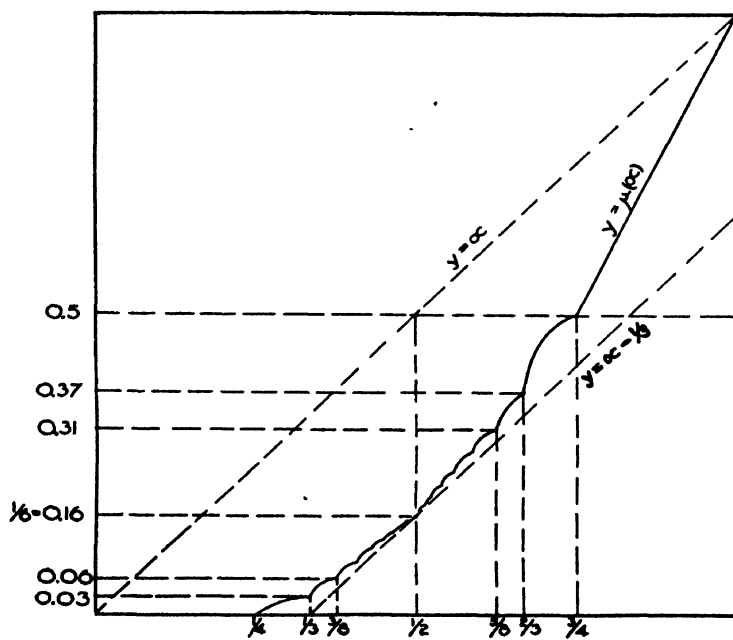


FIG. 7

The exact values of $\mu(\alpha)$ at $\alpha = \frac{1}{2}\left(1 \pm \frac{1}{n}\right)$, $n = 1, 2, 3, \dots$, are given by

$$\begin{aligned}\mu\left(\frac{1}{2}\left(1 - \frac{1}{n}\right)\right) &= \frac{(n-1)(n-2)}{6n^2} = \frac{1}{2}\left(1 - \frac{1}{n}\right) - \frac{1}{3} + \frac{1}{3n^2}. \\ \mu\left(\frac{1}{2}\left(1 + \frac{1}{n}\right)\right) &= \frac{(n+1)(n+2)}{6n^2} = \frac{1}{2}\left(1 + \frac{1}{n}\right) - \frac{1}{3} + \frac{1}{3n^2}.\end{aligned}$$

Thus the graph of $\mu(\alpha)$ lies entirely above the straight line $y = \alpha - \frac{1}{3}$, except at the point $\alpha = \frac{1}{2}$ where $\mu(\frac{1}{2}) = \frac{1}{6}$.

The author is indebted to Drs. W. Ambrose, D. Blackwell, R. H. Fox and P. R. Halmos for the conversations we have had in the course of this work. The values of $\mu(\alpha)$ for $\alpha = \frac{1}{2}\left(1 \pm \frac{1}{n}\right)$, $n = 1, 2, 3, \dots$ were obtained by R. H. Fox and P. R. Halmos.

II. Preliminary considerations

LEMMA 1.

$$(10) \quad \lambda(1 - \alpha) = 1 - 2\alpha + \lambda(\alpha),$$

$$(11) \quad \mu(1 - \alpha) = 1 - 2\alpha + \mu(\alpha).$$

PROOF. These are the direct consequences of the fact that $\varphi(x, y) \in \Phi$ implies $1 - \varphi(x, y) \in \Phi$, and that

$$(12) \quad A(1 - \varphi) = 1 - A(\varphi),$$

$$(13) \quad V(1 - \varphi) = 1 - 2A(\varphi) + V(\varphi),$$

which follow easily from the definitions of $A(\varphi)$ and $V(\varphi)$.

It is clear that the functions $\lambda(\alpha)$ and $\mu(\alpha)$ defined by (5), (6), (7), (8) and (9) satisfy (10) and (11). Hence it suffices to discuss either the case $0 \leq \alpha \leq \frac{1}{2}$ or the case $\frac{1}{2} \leq \alpha \leq 1$. This fact is needed in the following discussions.

DEFINITION 1. A real valued function $f(x)$ defined on the interval $I_x: 0 \leq x \leq 1$ is an *elementary function* if it is a finite linear combination of the characteristic functions of the intervals contained in I_x . (We do not care whether each of these intervals is closed or open, as we are only interested in elementary functions defined modulo null sets.) A set in the square $I_x \times I_y: 0 \leq x, y \leq 1$ is a *rectangular set*, if it is a direct product of two intervals each contained in I_x and I_y respectively. Finally, a real valued function defined on $I_x \times I_y$ is an *elementary function* if it is a finite linear combination of the characteristic functions of rectangular sets in $I_x \times I_y$.

Let Φ^0 be the subcollection of Φ consisting of all elementary functions $\varphi(x, y)$ in Φ . It is clear that in the definitions (3), (4) of $\lambda(\alpha)$ and $\mu(\alpha)$, we may replace the condition $\varphi \in \Phi$ by $\varphi \in \Phi^0$ and yet we obtain the same sup or inf. Hence, in order to prove our theorem, it suffices to show that $\mu(\alpha) \leq V(\varphi) \leq \lambda(\alpha)$

for any $\varphi(x, y) \in \Phi^0$ with $A(\varphi) = \alpha$, where $\lambda(\alpha)$ and $\mu(\alpha)$ are defined by (5), (6), (7), (8) and (9).

Let now $\varphi(x, y) \in \Phi^0$. We shall put

$$(14) \quad f_{\varphi}(x) = \int_0^1 \varphi(x, y) dy,$$

$$(15) \quad g_{\varphi}(y) = \int_0^1 \varphi(x, y) dx.$$

It is clear that $f_{\varphi}(x)$ and $g_{\varphi}(y)$ are elementary functions and we have

$$(16) \quad A(\varphi) = \int_0^1 f_{\varphi}(t) dt = \int_0^1 g_{\varphi}(t) dt,$$

$$(17) \quad V(\varphi) = \int_0^1 f_{\varphi}(t) g_{\varphi}(t) dt.$$

LEMMA 2. For any $\varphi(x, y) \in \Phi^0$, there exists a $\varphi'(x, y) \in \Phi^0$ such that $A(\varphi') = A(\varphi)$, $V(\varphi') = V(\varphi)$, and such that $f_{\varphi'}(x)$ is monotone non-increasing or monotone non-decreasing in x .

PROOF. Since $f_{\varphi}(x)$ is an elementary function, there exists a measure preserving transformation $x' = h(x)$ of the interval I_x onto itself, such that $f_{\varphi}(h(x))$ is monotone non-increasing or monotone non-decreasing. In fact, we can choose $h(x)$ as a permutation of subintervals of I_x . It is then clear that the function $\varphi'(x, y) = \varphi(h(x), y)$ satisfies all the conditions required in Lemma 2.

DEFINITION 2. A set $E \subset I_x \times I_y$ is an elementary set if it is a union of a finite number of rectangular sets. A set $E \subset I_x \times I_y$ is a corner set if $(x, y) \in E$, $0 \leq x' \leq x$, $0 \leq y' \leq y$ imply $(x', y') \in E$. Further, a set $E \subset I_x \times I_y$ is a corner* set if $(x, y) \in E$, $x \leq x' \leq 1$, $0 \leq y' \leq y$ imply $(x', y') \in E$.

LEMMA 3. Let $\varphi(x, y) \in \Phi^0$ be an elementary function such that $f_{\varphi}(x)$ is monotone non-increasing in x . Then there exists an elementary corner set $E \subset I_x \times I_y$ such that $A(\varphi_E) = A(\varphi)$, $V(\varphi_E) \geq V(\varphi)$.

PROOF. Let E be the set of all points $(x, y) \in I_x \times I_y$ such that $0 \leq y \leq f(x)$. It is clear that E is an elementary corner set satisfying

$$(18) \quad f_{\varphi_E}(t) = f_{\varphi}(t), \quad \text{for } 0 \leq t \leq 1.$$

From this follows easily that $A(\varphi_E) = A(\varphi)$. Moreover, it is easy to see that

$$(19) \quad G_{\varphi_E}(t) \geq G_{\varphi}(t), \quad \text{for } 0 \leq t \leq 1,$$

$$(20) \quad G_{\varphi_E}(0) = G_{\varphi}(0) (= 0), \quad G_{\varphi_E}(1) = G_{\varphi}(1) (= A(\varphi_E) = A(\varphi)),$$

where $G_{\varphi_E}(t)$ and $G_{\varphi}(t)$ are defined by

$$(21) \quad G_{\varphi_E}(t) = \int_0^t g_{\varphi_E}(s) ds, \quad G_{\varphi}(t) = \int_0^t g_{\varphi}(s) ds, \quad \text{for } 0 \leq t \leq 1$$

respectively.

Consequently,

$$\begin{aligned}
 (22) \quad V(\varphi_E) &= \int_0^1 f_{\varphi_E}(t) g_{\varphi_E}(t) dt \\
 &= [f_{\varphi_E}(t) G_{\varphi_E}(t)]_0^1 - \int_0^1 G_{\varphi_E}(t) df_{\varphi_E}(t) \\
 &\geq [f_{\varphi}(t) G_{\varphi}(t)]_0^1 - \int_0^1 G_{\varphi}(t) df_{\varphi}(t) \\
 &= \int_0^1 f_{\varphi}(t) g_{\varphi}(t) dt = V(\varphi),
 \end{aligned}$$

which proves Lemma 3.

In the same way we can prove

LEMMA 3'. Let $\varphi(x, y) \in \Phi^0$ be an elementary function such that $f_{\varphi}(x)$ is monotone non-decreasing in x . Then there exists an elementary corner* set $E \subset I_x \times I_y$ such that $A(\varphi_E) = A(\varphi)$, $V(\varphi_E) \leq V(\varphi)$.

We omit the proof.

DEFINITION 3. Let E be an elementary corner set in $I_x \times I_y$, and consider the graphs $y = f_{\varphi_E}(x)$ and $x = g_{\varphi_E}(y)$. These two graphs together will compose a polygonal line Γ_E , consisting only of horizontal and vertical segments, which connects the points $(0, 1)$ and $(1, 0)$. This polygonal line is called the *characteristic graph* of E . Similarly, we can define the characteristic graph of an elementary corner* set $E \subset I_x \times I_y$. This is a polygonal line connecting two points $(0, 0)$ and $(1, 1)$.

We shall divide our further arguments into two parts, namely, the discussion of $\lambda(\alpha)$ and that of $\mu(\alpha)$.

III. Discussion of $\lambda(\alpha)$

DEFINITION 4. A set $E \subset I_x \times I_y$ is *symmetric* if $(x, y) \in E$ implies $(y, x) \in E$.

LEMMA 4. For any elementary corner set $E \subset I_x \times I_y$, there exists a symmetric elementary corner set $E' \subset I_x \times I_y$ such that $A(\varphi_{E'}) = A(\varphi_E)$, $V(\varphi_{E'}) \geq V(\varphi_E)$.

PROOF. Let Γ_E be the characteristic graph of E . Γ_E has a unique intersection with the diagonal $x = y$ of the unit square $I_x \times I_y$. Let (ξ, ξ) be this point of intersection. Then the required set E' is defined as the set of all points $(x, y) \in I_x \times I_y$ satisfying one of the following three conditions:

$$(23) \quad 0 \leq x \leq \xi, \quad 0 \leq y \leq \xi,$$

$$(24) \quad 0 \leq x \leq \frac{1}{2}\{f_{\varphi_E}(y) + g_{\varphi_E}(y)\}, \quad \xi < y \leq 1,$$

$$(25) \quad \xi < x \leq 1, \quad 0 \leq y \leq \frac{1}{2}\{f_{\varphi_E}(x) + g_{\varphi_E}(y)\}.$$

It is clear that E' is a symmetric elementary corner set, and that $A(\varphi_{E'}) = A(\varphi_E)$. In order to prove that $V(\varphi_{E'}) \geq V(\varphi_E)$, we put

$$(26) \quad p(t) = f_{\sigma_E}(t), \quad q(t) = g_{\sigma_E}(t), \quad r(t) = f_{\sigma_E'}(t) = g_{\sigma_E'}(t), \quad \text{for } 0 \leq t \leq 1,$$

$$(27) \quad P(t) = \int_0^t p(s) ds, \quad Q(t) = \int_0^t q(s) ds, \quad R(t) = \int_0^t r(s) ds,$$

for $0 \leq t \leq 1$.

It is then easy to see that

$$(28) \quad 2R(t) \geq P(t) + Q(t), \quad \text{for } 0 \leq t \leq \xi,$$

$$(29) \quad 2r(t) = p(t) + q(t), \quad \text{for } \xi < t \leq 1.$$

Consequently,

$$\begin{aligned} \int_0^\xi \{r(t)^2 - p(t)q(t)\} dt &\geq \frac{1}{4} \int_0^\xi \{4r(t)^2 - (p(t) + q(t))^2\} dt \\ &= \frac{1}{4} \int_0^\xi \{2r(t) - p(t) - q(t)\} \{2r(t) + p(t) + q(t)\} dt \\ (30) \quad &= \frac{1}{4} [\{2R(t) - P(t) - Q(t)\} \{2r(t) + p(t) + q(t)\}]_\xi^\xi \\ &\quad - \frac{1}{4} \int_\xi^1 \{2R(t) - P(t) - Q(t)\} d\{2r(t) + p(t) + q(t)\} dt \geq 0 \end{aligned}$$

and

$$\begin{aligned} \int_\xi^1 \{r(t)^2 - p(t)q(t)\} dt &= \frac{1}{4} \int_\xi^1 \{(p(t) + q(t))^2 - p(t)q(t)\} dt \\ (31) \quad &= \frac{1}{4} \int_0^1 (p(t) - q(t))^2 dt \geq 0, \end{aligned}$$

which together imply

$$(32) \quad V(\varphi_{E'}) - V(\varphi_E) = \int_0^1 \{r(t)^2 - p(t)q(t)\} dt \geq 0.$$

The proof of Lemma 4 is completed.

Thus in order to discuss $\lambda(\alpha)$, it suffices to consider symmetric elementary corner sets $E \subset I_x \times I_y$.

Now let E be a symmetric elementary corner set in $I_x \times I_y$, and let Γ_E be its characteristic graph. Let us assume that Γ_E contains at least three segments above the diagonal $x = y$.¹ Let a, b, c be three consecutive segments of Γ_E lying above the diagonal $x = y$. We assume that a and c are horizontal, while b is vertical to the x -axis. (See Fig. 8.) We shall replace the part of Γ_E consisting of a, b, c by another system of segments a', b', c' as indicated in Fig. 8. Let us denote the new symmetric elementary corner set thus obtained by E' . (Of course, we make the same change below the diagonal, as indicated in Fig. 6,

¹ When we say that a segment (parallel to the x -axis or to the y -axis) lies above or below the diagonal $x = y$, one of the end points of the segment may lie on the diagonal $x = y$, while on the other hand, when we say that a segment lies *entirely* above or below the diagonal neither of the end points of the segment can lie on the diagonal $x = y$.

so as to make E' symmetric.) The y -coordinate of b' is so chosen that we have $A(\varphi_{E'}) = A(\varphi_E)$, and this condition is thus fulfilled if we have $\bar{a}/\bar{b}' = \bar{c}'/\bar{b} = \theta$, where θ is a real number satisfying $0 < \theta < 1$. We shall compare $V(\varphi_{E'})$ with $V(\varphi_E)$. A simple computation shows

$$(33) \quad V(\varphi_{E'}) - V(\varphi_E) = \theta(1 - \theta)(\bar{a} + \bar{c})\bar{b}(\bar{a} + \bar{c} - \bar{b}).$$

Hence, if $\bar{a} + \bar{c} > \bar{b}$, then by replacing a, b, c by a', b', c' we obtain a new symmetric elementary corner set E' such that $A(\varphi_{E'}) = A(\varphi_E)$, $V(\varphi_{E'}) \geq V(\varphi_E)$. If we interchange a, b, c with a', b', c' , then we immediately see that the same thing is true even if a and b are vertical, while b is horizontal to the x -axis.

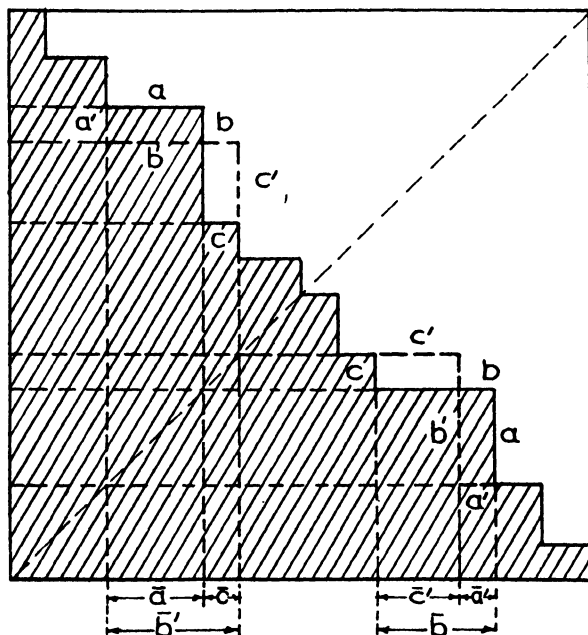


FIG. 8

Assume now that Γ_E contains at least four segments above the diagonal $x = y$. Let a, b, c, d be any four consecutive segments. We denote their respective lengths by $\bar{a}, \bar{b}, \bar{c}, \bar{d}$. It is then easy to see that at least one of the inequalities $\bar{a} + \bar{c} > \bar{b}$, $\bar{b} + \bar{d} > \bar{c}$ must hold. Hence, by replacing a, b, c or b, c, d by a suitable system a', b', c' or b', c', d' , we can always obtain a new symmetric elementary corner set E' for which $A(\varphi_{E'}) = A(\varphi_E)$, $V(\varphi_{E'}) \geq V(\varphi_E)$. Further, it is to be noticed that the number of segments lying above the diagonal $x = y$ in the characteristic graph of the new set E' is smaller than that of E exactly by two.

Thus, by iterating the same process, we shall finally reach a symmetric elementary corner set E^* whose characteristic graph Γ_{E^*} consists of at most three segments lying above the diagonal $x = y$, and such that $A(\varphi_{E^*}) = A(\varphi_E)$,

$V(\varphi_E) \geq V(\varphi_E)$. Consequently, in order to discuss $\lambda(\alpha)$ it suffices to consider the symmetric elementary sets E of the forms given in Figs. 1, 2, 9 and 10.

We shall discuss these cases separately.

(i) CASE OF FIG. 1. The condition $A(\varphi_E) = \alpha$ implies $\bar{a} = 1 - (1 - \alpha)^{\frac{1}{2}}$, $\bar{b} = (1 - \alpha)^{\frac{1}{2}}$. Consequently

$$(34) \quad V(\varphi_E) = 2\alpha - 1 + (1 - \alpha)^{\frac{1}{2}}.$$

(ii) CASE OF FIG. 2. The condition $A(\varphi_E) = \alpha$ implies $\bar{a} = 1 - \alpha^{\frac{1}{2}}$, $\bar{b} = \alpha^{\frac{1}{2}}$. Consequently,

$$(35) \quad V(\varphi_E) = \alpha^{\frac{1}{2}}.$$

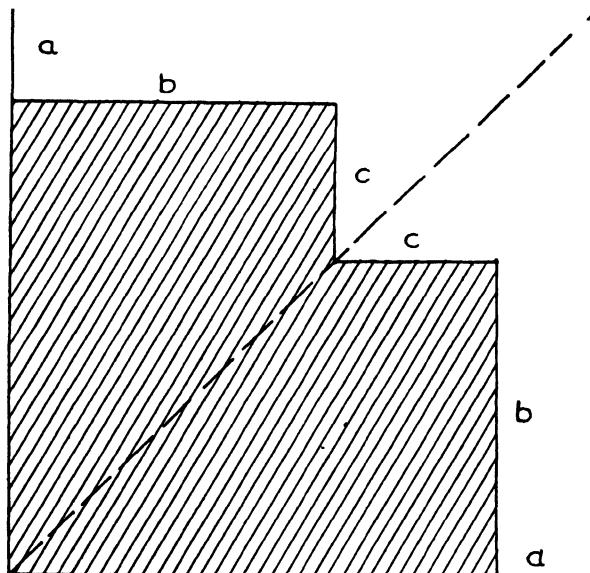


FIG. 9

(iii) CASE OF FIG. 9. A simple computation shows:

$$(36) \quad A(\varphi_E) = 2\bar{b}\bar{c} + \bar{b}^2 = \alpha,$$

$$(37) \quad V(\varphi_E) = \bar{b}(\bar{b} + \bar{c})^2 + \bar{b}^2\bar{c}.$$

Hence $c = (\alpha - \bar{b}^2)/2\bar{b}$. Putting this value in (37), we have

$$\begin{aligned} V(\varphi_E) &= \frac{1}{4\bar{b}} (\alpha^2 + 4\alpha\bar{b}^2 - \bar{b}^4) \\ &= \frac{1}{4\bar{b}} \{2\alpha^2 + 2\alpha\bar{b}^2 - (\alpha - \bar{b}^2)^2\} \\ (38) \quad &\leq \frac{1}{4\bar{b}} (2\alpha^2 + 2\alpha\bar{b}^2) = \frac{\alpha}{2} \left(\frac{\alpha}{\bar{b}} + \bar{b} \right) \\ &\leq \frac{\alpha}{2} \cdot 2 \left(\frac{\alpha}{\bar{b}} \cdot \bar{b} \right)^{\frac{1}{2}} = \alpha^{\frac{1}{2}}. \end{aligned}$$

(iv) CASE OF FIG. 10. A simple computation shows:

$$(39) \quad A(\varphi_E) = 1 - (2\bar{b}\bar{c} + \bar{b}^2) = \alpha,$$

$$(40) \quad \begin{aligned} V(\varphi_E) &= \bar{a} + \bar{c}(\bar{a} + \bar{c})^2 + \bar{a}^2\bar{b} \\ &= 2\alpha - 1 + \{\bar{b}(\bar{b} + \bar{c})^2 + \bar{b}^2\bar{c}\}. \end{aligned}$$

Consequently, by the result obtained above in the case of Fig. 7,

$$(41) \quad V(\varphi_E) \leq 2\alpha - 1 + (1 - \alpha)^{\frac{1}{2}}.$$

Summing up, we have thus proved that

$$(42) \quad V(\varphi) \leq \max(\alpha^{\frac{1}{2}}, 2\alpha - 1 + (1 - \alpha)^{\frac{1}{2}})$$

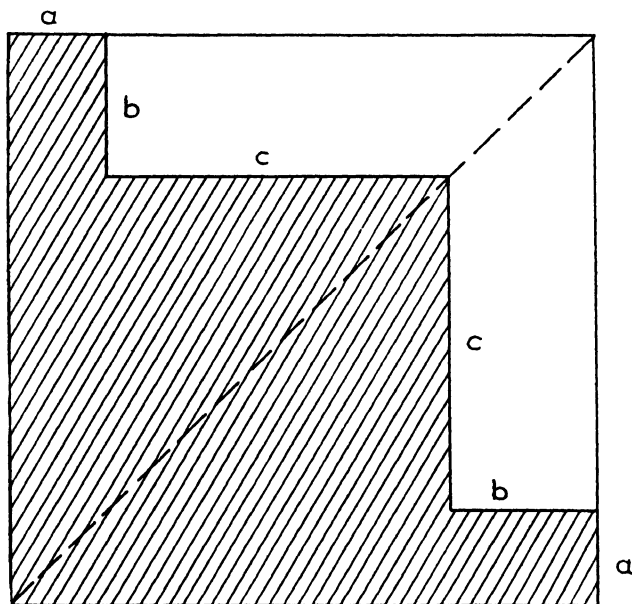


FIG. 10

for any $\varphi(x, y) \in \Phi$ with $A(\varphi) = \alpha$, $0 \leq \alpha \leq 1$, the equality holding for the symmetric elementary corner sets E of the forms given in Figs. 1 and 2. This completes the proof of our theorem for $\lambda(\alpha)$.

IV. Discussion of $\mu(\alpha)$

DEFINITION 5. Let E be an elementary corner* set in $I_x \times I_y$, and let Γ_E be its characteristic graph. E is a *special* elementary corner* set if there is no segment in Γ_E which lies *entirely*² above or below the diagonal $x = y$. For example, Fig. 11 shows a special elementary corner* set, while this is not the case in Fig. 12.

LEMMA 5. For any elementary corner* set $E \subset I_x \times I_y$, there exists a special elementary corner* set $E' \subset I_x \times I_y$, such that $A(\varphi_{E'}) = A(\varphi_E)$, $V(\varphi_{E'}) = V(\varphi_E)$.

² See footnote (1) on page 750.

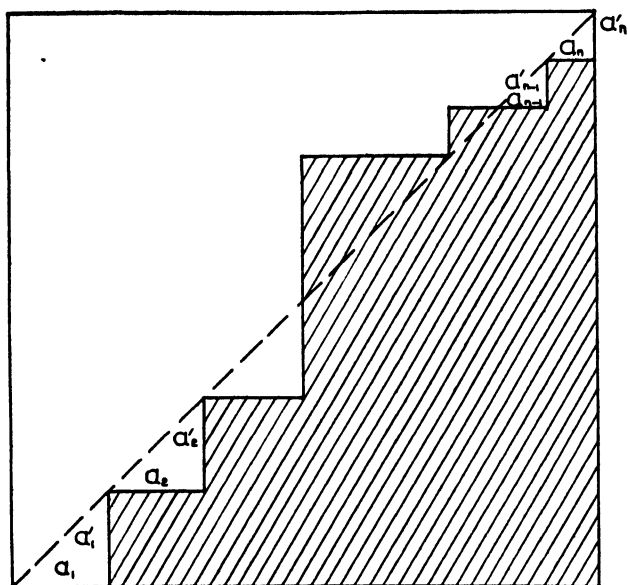


FIG. 11

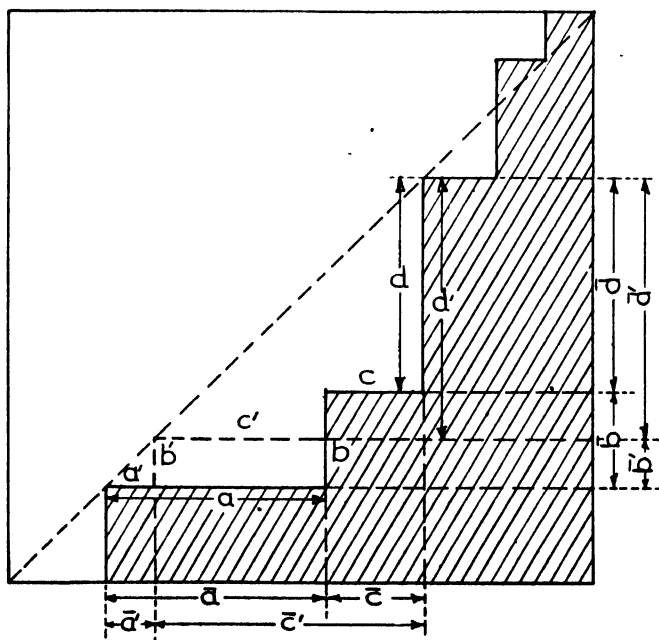


FIG. 12

PROOF. Let us assume that the given elementary corner* set E is of the form as given in Fig. 12. We shall replace the part of Γ_n consisting of a, b, c, d

by a', b', c', d' as indicated in Fig. 12. (Clearly we have $\bar{a} + \bar{c} = \bar{b} + \bar{d} = \bar{a}' + \bar{c}' = \bar{b}' + \bar{d}'$. Let us denote the new set thus obtained by E' . The condition $A(\varphi_{E'}) = A(\varphi_E)$ is fulfilled by taking $\bar{b}\bar{c} = \bar{b}'\bar{c}'$. Then a simple computation shows that $V(\varphi_{E'}) = V(\varphi_E)$.

Thus our lemma is proved if E is an elementary corner* set E of the form of Fig. 12. The analogous argument applies to the case when similar situation happens above the diagonal $x = y$. Finally, if there are more than two (and hence ≥ 4) consecutive segments or more than one pair of segments which lie entirely above or below the diagonal $x = y$, then we can iterate the same kind of operations, reducing the number of segments lying entirely above or below the diagonal $x = y$ exactly by two in each step, until we finally reach a special elementary corner* set E^* satisfying $A(\varphi_{E^*}) = A(\varphi_E)$, $V(\varphi_{E^*}) = V(\varphi_E)$. The proof of Lemma 5 is completed.

Thus, in order to discuss $\mu(\alpha)$ it suffices to consider special elementary corner* sets only.

Let now E be a special elementary corner* set in $I_x \times I_y$, as in Fig. 11. We denote the segments of its characteristic graph Γ_E successively by $a_1, a'_1, \dots, a_n, a'_n$ (see Fig. 11). Those a_i, a'_i which lie above the diagonal $x = y$ are denoted by b_j, b'_j , and those below the diagonal by c_k, c'_k . We have clearly, $\bar{a}_i = \bar{a}'_i, \bar{b}_j = \bar{b}'_j, \bar{c}_k = \bar{c}'_k$ and

$$(43) \quad \sum_i \bar{a}_i = \sum_j \bar{b}_j + \sum_k \bar{c}_k = 1.$$

Then a simple computation shows

$$\begin{aligned} (44) \quad A(\varphi_E) &= \frac{1}{2} \left\{ 1 + \sum_j \bar{b}_j^2 - \sum_k \bar{c}_k^2 \right\} = \alpha, \\ V(\varphi_E) &= \sum_{i < j < k} \bar{a}_i \bar{a}_j \bar{a}_k + \sum_j \bar{b}_j \\ &= \frac{1}{6} \left\{ \left(\sum_i \bar{a}_i \right)^3 - 3 \sum_i \bar{a}_i^2 \sum_i \bar{a}_i + 2 \sum_i \bar{a}_i^3 \right\} + \sum_j \bar{b}_j^2 \\ &= \frac{1}{6} \left\{ 1 - 3 \sum_i \bar{a}_i^2 + 2 \sum_i \bar{a}_i^3 \right\} + \sum_j \bar{b}_j^2 \\ (45) \quad &= \frac{1}{6} \left\{ 1 + 3 \left(\sum_j \bar{b}_j^2 - \sum_k \bar{c}_k^2 \right) + 2 \sum_i \bar{a}_i^3 \right\} \\ &= \frac{1}{6} \left\{ 1 + 3(2\alpha - 1) + 2 \sum_i \bar{a}_i^3 \right\} \\ &= \alpha - \frac{1}{3} + \frac{1}{3} \left\{ \sum_j \bar{b}_j^3 + \sum_k \bar{c}_k^3 \right\}. \end{aligned}$$

Thus our problem is transformed into the following one: *under the conditions (43) and*

$$(46) \quad \sum_j \bar{b}_j^3 - \sum_k \bar{c}_k^3 = 2\alpha - 1$$

to make

$$(47) \quad \sum_j \bar{b}_j^3 + \sum_k \bar{c}_k^3 = \omega = 3(V(\varphi_E) - \alpha) + 1$$

as small as possible, where $\bar{b}_j \geq 0$ and $\bar{c}_k \geq 0$ and there is no assumption on the number of \bar{b}_j and \bar{c}_k .

Let us consider the interval $0 \leq \alpha \leq \frac{1}{2}$. Then it is clear that we have only to consider the case when all $\bar{b}_j = 0$, and our problem is further reduced to the following one: Under the conditions:

$$(48) \quad \sum_k \bar{c}_k = 1,$$

$$(49) \quad \sum_k \bar{c}_k^2 = \beta \equiv 1 - 2\alpha > 0,$$

to make

$$(50) \quad \sum_k \bar{c}_k^3 = \omega = 3(V(\varphi_E) - \alpha) + 1$$

as small as possible, where $\bar{c}_k \geq 0$ and we have no assumption on the number n of c_k .

By Schwarz's inequality, we have $n\beta > 1$, and it is easy to see that for fixed n with $n\beta \geq 1$, the minimum value ω_n of ω is attained by

$$(51) \quad c_1 = \cdots = c_{n-1} = \frac{1}{n} \left(1 + \left(\frac{n\beta - 1}{n - 1} \right)^{\frac{1}{2}} \right),$$

$$(52) \quad c_n = \frac{1}{n} (1 - ((n\beta - 1)(n - 1))^{\frac{1}{2}})$$

and

$$(52) \quad \omega_n = \frac{1}{n^2} \left\{ (3n\beta - 2) - \frac{(n - 2)(n\beta - 1)^{\frac{1}{2}}}{(n - 1)^{\frac{1}{2}}} \right\}.$$

Hence, by (50) and (51), we finally have

$$\begin{aligned} V(\varphi_E) &= \alpha - \frac{1}{3} + \frac{1}{3n^2} \left\{ (3n(1 - 2\alpha) - 2) - \frac{(n - 2)(n(1 - 2\alpha) - 1)^{\frac{1}{2}}}{(n - 1)^{\frac{1}{2}}} \right\} \\ &= \frac{n - 2}{3n^2} \left\{ (3\alpha - 1)n + 1 - \frac{(n(1 - 2\alpha) - 1)^{\frac{1}{2}}}{(n - 1)^{\frac{1}{2}}} \right\} \end{aligned}$$

where φ_E is the characteristic function of a special elementary corner* set E of the form given in Fig. 3. It is easy to see that the expression (54) is a monotone increasing function of n for each given α for $n \geq (1 - 2\alpha)^{-1}$. Hence the smallest possible integer with $n\beta \equiv n(1 - 2\alpha) \geq 1$ gives the required value of $\mu(\alpha)$, or in other words, the equality (7) is true for $\alpha \geq \frac{1}{2} \left(1 - \frac{1}{n - 1} \right)$, $< \frac{1}{2} \left(1 - \frac{1}{n} \right)$.

Thus we have proved the formula (7) for $n = 2, 3, \dots$, i.e. for all α satisfying $0 \leq \alpha < \frac{1}{2}$. The formula (9) for $n = 2, 3, \dots$, or for $\frac{1}{2} < \alpha \leq 1$ then follows from this and from Lemma 1. Finally, the formula (8) for $\alpha = \frac{1}{2}$ follows from (7), (9) and from the fact that $\mu(\alpha)$ is a monotone non-decreasing function of α . This completes the discussion of $\mu(\alpha)$.

GROUP EXTENSIONS AND HOMOLOGY*

BY SAMUEL EILENBERG AND SAUNDERS MACLANE

(Received May 21, 1942)

CONTENTS

	<i>Page</i>
INTRODUCTION.....	758
CHAPTER I. TOPOLOGICAL GROUPS AND HOMOMORPHISMS.....	760
1. Topological spaces.....	760
2. Topological groups.....	761
3. The group of homomorphisms.....	762
4. Free groups and their factor groups.....	763
5. Closures and extendable homomorphisms.....	765
CHAPTER II. GROUP EXTENSIONS.....	766
6. Definition of extensions.....	767
7. Factor sets for extensions.....	767
8. The group of extensions.....	770
9. Group extensions and generators.....	771
10. The connection between homomorphisms and factor sets.....	771
11. Applications.....	774
12. Natural homomorphisms.....	777
CHAPTER III. EXTENSIONS OF SPECIAL GROUPS.....	778
13. Characters.....	778
14. Modular traces.....	780
15. Extensions of compact groups.....	782
16. Two lemmas on homomorphisms.....	784
17. Extensions of integers.....	785
18. Tensor products.....	787
CHAPTER IV. DIRECT AND INVERSE SYSTEMS.....	789
19. Direct systems of groups.....	789
20. Inverse systems of groups.....	789
21. Inverse systems of homomorphisms.....	791
22. Inverse systems of group extensions.....	791
23. Contracted extensions.....	793
24. The group Ext^*	794
25. Relation to tensor products.....	797
CHAPTER V. ABSTRACT COMPLEXES.....	798
26. Complexes.....	799
27. Homology and cohomology groups.....	800
28. Topology in the homology groups.....	802

* Presented to the American Mathematical Society, September 4 and December 31, 1941. Part of the results was published in a preliminary report [5] and also in an appendix to Lefschetz [7]. The numbers in brackets refer to the bibliography at the end of the paper.

29. The Kronecker index.....	803
30. Construction of homomorphisms.....	804
31. Study of A^q	806
32. Computation of the homology groups.....	808
33. Computation of the cohomology groups.....	809
34. The groups H_i^q	811
35. Universal coefficients.....	812
36. Closure finite complexes.....	813
CHAPTER VI. TOPOLOGICAL SPACES.....	814
37. Chain transformations.....	814
38. Naturality.....	815
39. Čech's homology groups.....	816
40. Formulas for a general space.....	818
41. The case $q = 0$	820
42. Fundamental complexes.....	820
43. Relations between a space and its fundamental complex.....	821
44. Formulas for a compact metric space.....	823
45. Regular cycles.....	824
APPENDIX A. COEFFICIENT GROUPS WITH OPERATORS.....	825
APPENDIX B. SOLENOIDS.....	829
BIBLIOGRAPHY.....	831

INTRODUCTION

In 1937 the following problem was formulated by Borsuk and Eilenberg: Given a solenoid¹ Σ in the three sphere S^3 , how many homotopy classes of continuous mappings $f(S^3 - \Sigma) \subset S^2$ are there? In 1939 Eilenberg proved ([4], p. 251) that the homotopy classes in question are in a 1-1-correspondence with the elements of the one-dimensional homology group $H^1(K, I) = Z^1(K, I)/B^1(K, I)$, where K is any representation of $S^3 - \Sigma$ as a complex, $Z^1(K, I)$ is the group of infinite 1-cycles in K with the additive group I of integers as coefficients and $B^1(K, I)$ is the subgroup of bounding cycles. This homology group is generally much "larger" than the conventional homology group $H_i(K, I) = Z^i/B^i$ where $B^i(K, I)$ is the group of cycles that bound on every finite portion of K ; with an appropriate topology in the group Z^1 , B^1 turns out to be exactly the closure of B^1 .

At this point the investigation was taken up by Steenrod [10]. By using "regular cycles" he computed the groups $H^1(S^3 - \Sigma)$ for the various solenoids Σ . The groups are uncountable and of a rather complicated nature.²

This paper originated from an accidental observation that the groups obtained by Steenrod were identical with some groups that occur in the purely algebraic theory of *extensions of groups*. An abelian group E is called an ex-

¹ For the definition see Appendix B below.

² A popular exposition of Steenrod's results can be found in his article in *Lectures in Topology*, Ann Arbor, University of Michigan Press, 1941, pp. 43-55.

tension of the group G by the group H if $G \subset E$ and $H = E/G$. With a proper definition of equivalence and addition, the extensions of G by H themselves form an abelian group $\text{Ext } \{G, H\}$. It turns out that $H^1(S^3 - \Sigma, I)$ is isomorphic with $\text{Ext } \{I, \Sigma^*\}$ where Σ^* is a properly chosen subgroup of the group of rational numbers.³

The thesis of this paper is that the theory of group extensions forms a natural and powerful tool in the study of homologies in infinite complexes and topological spaces. Even in the simple and familiar case of finite complexes the results obtained are finer than the existing ones.

Our fundamental theorem concerns the homology groups of a star finite complex K . Let $H^q(G)$ denote the homology group of infinite cycles with coefficients in an arbitrary topological group G . We obtain an explicit expression for $H^q(G)$ in terms of G and the cohomology groups \mathcal{K}_q of *finite* cocycles with integral coefficients. (\mathcal{K}_q is the factor group $\mathcal{Z}_q/\mathcal{B}_q$ of cocycles modulo coboundaries). This expression is

$$H^q(G) = \text{Hom } \{\mathcal{K}_q, G\} \times \text{Hom } \{\mathcal{B}_{q+1}, G\} / \text{Hom } \{\mathcal{Z}_{q+1} \mid \mathcal{B}_{q+1}, G\}.$$

Here $\text{Hom } \{H, G\}$ stands for the (topological) group of all homomorphisms of H into G , while $\text{Hom } \{\mathcal{Z}_{q+1} \mid \mathcal{B}_{q+1}, G\}$ denotes the group of those homomorphisms of \mathcal{B}_{q+1} into G which can be extended to homomorphisms of \mathcal{Z}_{q+1} into G . The factor group on the right in this expression appears to depend on the groups \mathcal{B}_{q+1} and \mathcal{Z}_{q+1} , but actually depends only on the cohomology group $\mathcal{K}_{q+1} = \mathcal{Z}_{q+1}/\mathcal{B}_{q+1}$. In fact this factor group can best be interpreted as the group "Ext" of group extensions of G by \mathcal{K}_{q+1} . The fundamental theorem then has the form

$$H^q(G) = \text{Hom } \{\mathcal{K}_q, G\} \times \text{Ext } \{G, \mathcal{K}_{q+1}\}.$$

The paper is self contained as far as possible, both in algebraic and topological respects. The first four chapters below develop the requisite group-theoretical notions. Chapter I discusses the groups of homomorphisms involved in the above formula, while Chapter II introduces the group of group extensions, and proves the fundamental theorem relating this group to groups of homomorphisms. This fundamental theorem is essentially a formulation of the known fact that a group extension of G by H can be described either by generators of H (and hence by homomorphisms) or by certain "factor sets." Chapter III analyzes the group $\text{Ext } \{G, H\}$ for some special cases of G . Chapter IV introduces some additional groups, closely related to Ext , which arise as inverse limit groups in the treatment of homologies of topological spaces.

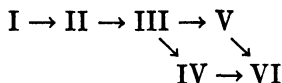
The last two chapters analyze homology groups. Chapter V treats the case of a complex, and proves the fundamental theorem quoted above, as well as parallel theorems for some of the other homology groups of a complex. Chapter

³ More precisely Σ^* is the character group of Σ . The detailed treatment appears in Appendix B below.

VI obtains analogous theorems for the Čech homology groups of a topological space.

Appendix A discusses the case when G is a group with operators. Appendix B contains a computation of the group $\text{Ext}\{I, \Sigma^*\}$ mentioned above.

Each chapter is preceded by a brief outline. The chapters are related as in the following diagram:



Almost all of V can be read directly after I and II, and a major portion after I alone.

Chapters V and VI are strongly influenced by S. Lefschetz's recent book "Algebraic Topology" [7], that the authors had the privilege of reading in manuscript.

CHAPTER I. TOPOLOGICAL GROUPS AND HOMOMORPHISMS

After a certain preliminary definitions, this chapter introduces the basic group $\text{Hom}\{R, G\}$ of homomorphisms. In the case when R is a subgroup of a free group, we require two subgroups of "extendable" homomorphisms. The topology of these subgroups is investigated when the "coefficient group" G is itself topological.

1. Topological spaces

A set X is called a *space* if there is given a family of subsets of X , called *open sets*, such that

- (1.1) X and the void set are open,
- (1.2) the union of any number of open sets is open,
- (1.3) the intersection of two open sets is open.

Complements of open sets are called *closed*. X is called a *Hausdorff space* if in addition

- (1.4) every two distinct points are contained respectively in two disjoint open sets.
- X is called a *compact* (= bcompact) space if
- (1.5) every covering of X by open sets contains a finite subcovering.

A space X is *discrete* if every set in X is open.

The intersection of an open set of a space X with a subset A of X will be called *open in A* . With this convention A becomes a space.

Let X and Y be spaces and $x \rightarrow f(x) = y$ a mapping of X into a subset of Y . The mapping f is *continuous* if for every open set $U \subset Y$ the set $f^{-1}(U)$ is open (in X). The mapping f is *open* if for every open set $U \subset X$ the set $f(U)$ is open (in Y). A well known result is

LEMMA 1.1. *If f is a continuous mapping of a compact space X into a Hausdorff space Y , then $f(X)$ is closed in Y .*

A *product space* $\prod_{\alpha} X_{\alpha}$ of a given collection $\{X_{\alpha}\}$ of spaces X_{α} is defined as

the space whose points are all collections $\{x_\alpha\}$, $x_\alpha \in X_\alpha$ and in which open sets are unions of sets of the form $\prod_\alpha U_\alpha$, where U_α is an open subset of X_α and $U_\alpha = X_\alpha$ except for a finite number of indices α .⁴ It is known that $\prod X_\alpha$ is a Hausdorff or compact space if and only if for every α the space X_α is a Hausdorff or compact space.⁵

Let Λ be a set of elements and X be a space. We consider the set X^Λ of all functions with arguments in Λ and values in X . The set X^Λ is clearly in a 1-1 correspondence with the product $\prod X_\lambda$ where $\lambda \in \Lambda$ and $X_\lambda = X$. Hence we may consider X^Λ as a space.

2. Topological groups

Only abelian groups (written additively) will be considered.

A group G will be called a *generalized topological group* if G is a space in which the group composition (as a mapping $G \times G \rightarrow G$) and the group inverse (as a mapping $G \rightarrow G$) are continuous.

If G , considered as a space, is a Hausdorff space, then G will be called a *topological group*.⁶ Similarly, if G is compact as a space we shall say that G is a *compact group*.

A subgroup of a (generalized) topological group is a (generalized) topological group. A closed subgroup of a compact group is compact.

LEMMA 2.1. *In a generalized topological group G the following properties are equivalent:*

- (a) every point of G is a closed set,
- (b) the zero element of G is a closed set,
- (c) G is a topological group.⁷

The factor group $H = G/G_1$ of a generalized topological group G modulo a subgroup G_1 is the group of all cosets $g + G_1$ of G_1 in G . The correspondence $\varphi(G) = H$ carrying each $g \in G$ into its coset $\varphi g = g + G_1$ in H is the "natural" mapping of G on H . We introduce a topology in H by calling a set $U \subset H$ open if and only if $\varphi^{-1}(U)$ is open in G . It can be shown that this topology is the only one under which φ will be both open and continuous.

LEMMA 2.2. *If G is a generalized topological group and G_1 is an arbitrary subgroup of G , then the factor group $H = G/G_1$ is a generalized topological group; it is a topological group if and only if G_1 is a closed subgroup of G . If G is compact, then G/G_1 is compact.*

LEMMA 2.3. *The closure $\bar{0}$ of the zero element of a generalized topological group is a closed subgroup of G . Its factor group $G/\bar{0}$ is the "largest" factor group of G which is a topological group.*

The preceding two statements show the utility of the study of generalized

⁴ If $\{\alpha\} = 1, 2, \dots, n$ we also use the symbol $X_1 \times X_2 \times \dots \times X_n$ for the product space.

⁵ See C. Chevalley and O. Frink, *Bulletin Amer. Math. Soc.* 47 (1941), pp. 612-614.

⁶ G is then a topological group in the sense of Pontrjagin [8].

⁷ To prove that a) implies c) one first proves that each neighborhood of g contains the closure of a neighborhood of g , as in Pontrjagin [8], p. 43, proposition F.

topological groups. Several times in the sequel we need to consider an isomorphism

$$(2.1) \quad G_1/H_1 \cong G_2/H_2$$

where the G_i are topological groups, while the H_i are not closed, so that G_i/H_i are only generalized topological groups. However, if we are able to prove that the isomorphism (2.1) is continuous in both directions in the "generalized" topology of the groups G_i/H_i , we obtain as a corollary the bicontinuous isomorphism of the topological groups G_i/\bar{H}_i .

If $\{G_\alpha\}$ is a collection of generalized topological groups the direct product $\prod_\alpha G_\alpha$ is a generalized topological group, provided we define the sum $\{g'_\alpha\} = \{g_\alpha\} + \{g''_\alpha\}$ by setting $g'_\alpha = g_\alpha + g''_\alpha$ for every α . Similarly, if Λ is any set and G is a generalized topological group, then the set G^Λ of all mappings of Λ into G is a generalized topological group. It follows from the results quoted in §1 that $\prod_\alpha G_\alpha$ and G^Λ are topological or compact groups if and only if the groups G_α and G are all topological or compact, respectively.

3. The group of homomorphisms

Let G and H be generalized topological groups. A homomorphism θ of H into G is a continuous function $\theta(h)$ defined for all $h \in H$ with values in G , such that $\theta(h_1 + h_2) = \theta(h_1) + \theta(h_2)$. For instance, the natural mapping of a group into one of its factor groups is a homomorphism. If θ_1 and θ_2 are two homomorphisms their sum $\theta_1 + \theta_2$, defined by

$$(\theta_1 + \theta_2)(h) = \theta_1(h) + \theta_2(h), \quad (\text{all } h \text{ in } H)$$

is also a homomorphism. Under this addition, the set of all homomorphisms θ of H into G constitutes a group, which we denote by $\text{Hom } \{H, G\}$:

$$(3.1) \quad \text{Hom } \{H, G\} = \{\text{all homomorphisms } \theta \text{ of } H \text{ into } G\}.$$

To introduce a (generalized) topology in $\text{Hom } \{H, G\}$, take any compact subset X of H and any open subset V of G with $0 \in V$ and consider the set $U(X, V)$ of all θ with $\theta(X) \subset V$. In the usual sense ([8], p. 55) these sets $U(X, V)$ constitute a complete set of neighborhoods of 0 in $\text{Hom } \{H, G\}$, and are used to define the topology of $\text{Hom } \{H, G\}$.⁸

If H is discrete, the compact subsets X of H are just the finite ones. In this case $\text{Hom } \{H, G\}$ is a subgroup of the group G^H with the topology as defined in §2.

LEMMA 3.1. *If G is a topological group and H is discrete, then $\text{Hom } \{H, G\}$ is a closed subgroup of the group G^H of all mappings of H into G .*

PROOF. Let $\phi_0 \in G^H$ be a mapping of H into G that is not a homomorphism. There are then elements h_1, h_2, h_3 in H such that $h_1 + h_2 = h_3$ and $\phi_0(h_1) + \phi_0(h_2) \neq \phi_0(h_3)$. Since G is a Hausdorff space and the group composi-

⁸ This is the general definition stated by Weil [11], p. 99, and Lefschetz [7], Ch. II.

tion is continuous there are in G three open sets U_1, U_2, U_3 containing $\phi_0(h_1), \phi_0(h_2)$, and $\phi_0(h_3)$, respectively, such that⁹ $(U_1 + U_2) \cap U_3 = 0$. Consequently the open subset U of G^H consisting of the mappings ϕ such that $\phi(h_1) \in U_1, \phi(h_2) \in U_2$, and $\phi(h_3) \in U_3$ has no elements in common with $\text{Hom } \{H, G\}$. Hence $\text{Hom } \{H, G\}$ is closed.

COROLLARY 3.2. *If H is discrete and G is a topological (and compact) group, then $\text{Hom } \{H, G\}$ is a topological (and compact) group.*

Note that the topology of $\text{Hom } \{H, G\}$ may not be discrete even though H and G both have discrete topologies. Observe also that if H is discrete, an alteration in the topology of G may alter the topology of $\text{Hom } \{H, G\}$ but not its algebraic structure. However, if H carries a non-discrete topology, an alteration in the topology of either H or G may alter the algebraic structure of $\text{Hom } \{H, G\}$, in that continuous homomorphisms may cease to be continuous, or vice versa.

If H is compact, we can take H itself to be the compact set X used in the definition of the topology in $\text{Hom } \{H, G\}$. Consequently, given any open set V in G containing 0, the homomorphisms θ , such that $\theta(H) \subset V$, constitute an open set. Hence if V can be picked so as not to contain any subgroups but 0, we see that $\text{Hom } \{H, G\}$ is discrete.

Subgroups and factor groups of H will correspond respectively to factor groups and subgroups of $\text{Hom } \{H, G\}$, as stated in the following lemmas.

LEMMA 3.3. *If H/H_1 is a factor group of the discrete group H , then $\text{Hom } \{H/H_1, G\}$ is (bicontinuously) isomorphic to that subgroup of $\text{Hom } \{H, G\}$ which consists of the homomorphisms θ mapping every element of H_1 into zero.*

The proof is readily given by observing that each homomorphism θ with $\theta(H_1) = 0$ maps each coset of H_1 into a single element of G , so induces a homomorphism θ' of H/H_1 . The continuity of the isomorphism $\theta \rightarrow \theta'$ can be established, as always for isomorphisms between groups, by showing continuity at $\theta = 0$. ([8], p. 63).

LEMMA 3.4. *If L is a subgroup of H , then each homomorphism θ of H into G induces a homomorphism $\theta' = \theta|L$ of L into G . The correspondence $\theta \rightarrow \theta'$ is a (continuous) homomorphism of $\text{Hom } \{H, G\}$ into $\text{Hom } \{L, G\}$. If L is a direct factor of H , this correspondence maps $\text{Hom } \{H, G\}$ onto $\text{Hom } \{L, G\}$.*

4. Free groups and their factor groups

The homology groups will be interpreted later as certain groups of homomorphisms of "free" groups, which we now define. If the elements z_α of a discrete group F are such that every element of F can be represented uniquely as a finite sum $\sum n_\alpha z_\alpha$ with integral coefficients n_α , F is said to be a *free abelian group* with generators (or basis elements) $\{z_\alpha\}$. The number of generators may be infinite. A free group can be constructed with any assigned set of symbols as basis elements.

⁹ $U_1 + U_2$ is the set of all sums $g_1 + g_2$, with $g_i \in U_i$. The symbol \cap stands for the set-theoretic intersection.

LEMMA 4.1. *Every proper subgroup of a free group is free.*

For the denumerable case, this is proved by Čech [3]; a general proof is given in Lefschetz [7] (II, (10.1)).

Any discrete group H can be represented as a homomorphic image of a free group. Specifically, if we choose any set of elements t_α in H which together generate all of H , and if we then construct a free group F with generators z_α in 1-1 correspondence $z_\alpha \leftrightarrow t_\alpha$ with the given t 's, the correspondence $\sum n_\alpha z_\alpha \rightarrow \sum n_\alpha t_\alpha$ will map the free group F homomorphically onto the given group H . If the kernel of this homomorphism¹⁰ is R , H may be represented as the factor group $H = F/R$. R is essentially the group of "relations" on the generators t_α of H .

Given $R \subset F$, each homomorphism ϕ of F into G induces a homomorphism $\theta = \phi|_R$ of the subgroup R into G , and the homomorphisms so induced form a subgroup of $\text{Hom}\{R, G\}$, denoted as

$$(4.1) \quad \text{Hom}\{F|_R, G\} = [\text{all } \theta = \phi|_R, \text{ for } \phi \in \text{Hom}\{F, G\}].$$

Alternatively, the elements of this subgroup can be described as those homomorphisms θ of R into G which can be extended (in at least one way) to homomorphisms of F into G .

A similar, but lighter, restriction may be imposed as follows: Given $\theta \in \text{Hom}\{R, G\}$, require that for every subgroup $F_0 \supset R$ of F for which F_0/R is finite there exist an extension of θ to a homomorphism of F_0 into G . The θ 's meeting this requirement also constitute a subgroup,

$$(4.2) \quad \text{Hom}_f\{R, G; F\} = [\text{all } \theta \in \text{Hom}\{F_0|_R, G\} \text{ for every finite } F_0/R].$$

These two subgroups,

$$\text{Hom}\{F|_R, G\} \subset \text{Hom}_f\{R, G; F\} \subset \text{Hom}\{R, G\},$$

are important because the corresponding factor groups in $\text{Hom}\{R, G\}$ are invariants of the group $H = F/R$, in that they do not depend on the particular free group F chosen to represent H . This fact may be stated as follows.

THEOREM 4.2. *If H is isomorphic to two factor groups F/R and F'/R' of free groups F and F' , then*

$$(4.3) \quad \text{Hom}\{R, G\}/\text{Hom}\{F|_R, G\} \cong \text{Hom}\{R', G\}/\text{Hom}\{F'|_{R'}, G\},$$

the isomorphism being both algebraic and topological. The same result holds for the factor groups

$$(4.4) \quad \text{Hom}\{R, G\}/\text{Hom}_f\{R, G; F\}, \quad \text{Hom}_f\{R, G; F\}/\text{Hom}\{F|_R, G\}.$$

This theorem is a corollary of a result to be established in Chapter II, as Theorem 10.1. It can also be proved directly, by appeal to the following lemma, which we state without proof.

¹⁰ The kernel of a homomorphism θ of a group H is the set of all elements $h \in H$ with $\theta(h) = 0$.

LEMMA 4.3. *Let $F/R = E/G$, where $F \supset R$ is a free group and $E \supset G$ is any other group. There exists a homomorphism ϕ of F into E such that, in the given identification of cosets of G with cosets of R ,*

$$(4.5) \quad \phi(x) + G = x + R, \quad \text{for all } x \in F.$$

Any other $\phi^ \in \text{Hom} \{F, E\}$ with this property (4.5) has the form $\phi^* = \phi + \beta$, for some $\beta \in \text{Hom} \{F, G\}$. Conversely, given ϕ with the property (4.5) any such $\phi^* = \phi + \beta$ has the same property.*

Although a given group H can be represented in many ways as a factor group $H = F/R$ of a free group, there is a "natural" such representation, in which F is the additive group F_H of the (integral) group ring of H . Specifically, given H , we choose for each $h \in H$ a symbol z_h and construct a free group F_H generated by the symbols z_h . The correspondence $z_h \rightarrow h$ induces a homomorphism of F_H on H . Let R_H denote the kernel of this homomorphism. The factor group (4.3) of the Theorem can then be described invariantly in terms of H and G as the group

$$\text{Hom} \{R_H, G\} / \text{Hom} \{F_H \mid R_H, G\}.$$

The same remark applies to the factor groups of (4.4). It would be possible to use the groups so described as substitutes for the group of group extensions to be introduced in Chapter II.

5. Closures and extendable homomorphisms

If G is topological, we wish to examine the closures of the groups $\text{Hom} \{F \mid R, G\}$ and Hom_f in the topological group $\text{Hom} \{R, G\}$. A preliminary is a characterization of the subgroup Hom_f .

LEMMA 5.1. *A homomorphism θ of $\text{Hom} \{R, G\}$ lies in $\text{Hom}_f \{R, G; F\}$ if and only if for each element t in F with a multiple mt in R there exists $h \in G$ with $\theta(mt) = mh$.*

PROOF. Let F_t be the subgroup of F generated by t and R . If $mt \in R$ for $m \neq 0$, F_t/R is finite and cyclic, so that $\theta \in \text{Hom}_f$ is extendable to F_t . Hence the condition stated on $\theta(mt)$ is necessary. Conversely, for any given group $F_0 \subset F$ with F_0/R finite we can write F_0/R as a direct product of cyclic groups. By applying the given condition on θ to each of these cyclic groups, we find an extension of θ to F_0 , as required.

Another characterization of Hom_f can be found; the proof is similar:

LEMMA 5.2. *A homomorphism θ of $\text{Hom} \{R, G\}$ lies in $\text{Hom}_f \{R, G; F\}$ if and only if θ can be extended to a homomorphism (into G) of each subgroup F_0 of F which contains R and for which the factor group F_0/R has a finite number of generators.*

We now consider the topology on $\text{Hom} \{R, G\}$.

LEMMA 5.3. *If G and hence $\text{Hom} \{R, G\}$ are generalized topological groups, $\text{Hom}_f \{R, G; F\}$ is contained in the closure of $\text{Hom} \{F \mid R, G\}$, or*

$$\text{Hom} \{F | R, G\} \subset \text{Hom}_f \{R, G; F\} \subset \overline{\text{Hom}} \{F | R, G\} \subset \text{Hom} \{R, G\}.$$

PROOF. Let θ_0 be in $\text{Hom}_f \{R, G; F\}$, while U is any open set of $\text{Hom} \{R, G\}$ containing θ_0 . Since F is discrete, the definition of the topology in $\text{Hom} \{R, G\}$ implies that there is a finite set of elements r_1, \dots, r_n of R such that U contains all θ for which each $\theta(r_i) = \theta_0(r_i)$. The elements r_i are all contained in a subgroup F_0 of F generated by a finite number of the given independent generators of the free group F . Since $\theta_0 \in \text{Hom}_f$, θ_0 has an extension θ' to the group generated by F_0 and R (Lemma 5.2). Introduce a new homomorphism θ^* of F by setting $\theta^*(z_\alpha) = \theta'(z_\alpha)$ for each generator z_α of F_0 , $\theta^*(z_\alpha) = 0$ otherwise. This θ^* induces a homomorphism θ of R , which agrees with θ_0 on the original elements r_1, \dots, r_n and which is by construction an element of $\text{Hom} \{F | R, G\}$. In other words, the arbitrary neighborhood U of θ_0 does contain a homomorphism $\theta \in \text{Hom} \{F | R, G\}$. This proves the lemma.

LEMMA 5.4. *If G is a compact topological group, $\text{Hom} \{F | R, G\}$ is a closed sub-group of $\text{Hom} \{R, G\}$, and hence $\text{Hom} \{F | R, G\} = \text{Hom}_f \{R, G; F\}$.*

PROOF. By Corollary 3.2, both the groups $\text{Hom} \{R, G\}$ and $\text{Hom} \{F, G\}$ are compact and topological. The second of these groups is mapped homomorphically onto $\text{Hom} \{F | R, G\}$ by the continuous correspondence $\theta \mapsto \theta | R$ of Lemma 3.4. Therefore, by Lemma 1.1, the image $\text{Hom} \{F | R, G\}$ is closed.

For any integer m , let mG be the subgroup of all elements of the form mg , with g in G . A condition for the closure of Hom_f may be stated in terms of these subgroups.

LEMMA 5.5. *If G is a generalized topological group, then $\text{Hom}_f \{R, G; F\}$ is closed in $\text{Hom} \{R, G\}$ whenever every subgroup mG of G is closed in G , for $m = 2, 3, \dots$ ¹¹.*

PROOF. Let θ be a homomorphism in the closure of $\text{Hom}_f \{R, G; F\}$. Consider an arbitrary t in F such that $mt \in R$. By Lemma 5.1 and the given condition on G it will suffice to prove that $\theta(mt) \in mG$. Let V be any open set containing 0 in G . By the definition of the topology in $\text{Hom} \{R, G\}$, there exists for θ in the closure of Hom_f an element θ' in Hom_f itself, such that $\theta'(mt) - \theta(mt) \in V$. But $\theta'(mt)$ is in mG , so that the arbitrary open set $V + \theta(mt)$ does contain an element of mG . This proves $\theta(mt)$ in mG , as required.

An examination of this proof shows that the given condition on G can be somewhat weakened. It suffices to require that the subgroup mG be closed in G for every integer m which is the order of an element of F/R . The same remark will apply in various subsequent cases when this condition on G is used.

CHAPTER II. GROUP EXTENSIONS

This chapter introduces the basic group $\text{Ext} \{G, H\}$ of all group extensions of G by H , and its subgroup $\text{Ext}_f \{G, H\}$ of all extensions which are "finitely trivial"

¹¹ If every subgroup mG is closed in G , Steenrod [9] and Lefschetz [7] say that G has the "division closure property."

(§8). Each individual group extension can be described either by a suitable "factor set" (§7) or by a certain homomorphism. The equivalence of these two representations is the fundamental theorem of this chapter (Theorem 10.1); it gives an expression of $\text{Ext } \{G, H\}$ as one of the factor-homomorphism groups already considered in Chapter I. This fundamental theorem, which is implicit in previous algebraic work on group extensions, is of independent algebraic interest. The chapter closes with a proof that the representation of $\text{Ext } \{G, H\}$ by homomorphisms is a "natural" one (§12). This conclusion is needed for the subsequent limiting process, which is used in defining the Čech homology groups.

6. Definition of extensions

A group E having G as subgroup and $H = E/G$ as the corresponding factor group is said to be an "extension" of G by H . More explicitly, if the groups G and H are given, a *group extension* of G by H is a pair (E, β) , where E is a group containing G and β is a homomorphism of E onto H under which exactly the elements of G are mapped into $0 \in H$.¹² Such a β induces an isomorphism of E/G to H . For given G and H , two extensions (E_1, β_1) and (E_2, β_2) are regarded as *equivalent* if and only if there is an isomorphism ω of E_1 to E_2 which leaves elements of G and cosets of H fixed. In other words, the isomorphism ω of E_1 to E_2 must have $\omega g = g$ for $g \in G$ and $\beta_2 \omega x = \beta_1 x$ for $x \in E_1$. We regard equivalent extensions as identical, and so study the equivalence classes of extensions of G by H . It will appear that these equivalence classes are themselves the elements of a group.

For given G and H , the direct product $G \times H$ has the "natural" homomorphism $(g, h) \rightarrow h$ onto H , and so can be regarded as an extension of G by H . Any extension (E, β) equivalent to this direct product (with its natural homomorphism) is said to be a *trivial* extension of G by H .

7. Factor sets for extensions

A given extension (E, β) of G by H can be described in terms of representatives for elements of H . To each h in H select in E a representative $u(h)$, such that $\beta(u(h)) = h$. Every element of E lies in some coset h , so has the form $g + u(h)$ for g in G . The sum of any two representatives $u(h)$ and $u(k)$ will lie in the same coset, modulo G , as does the representative of the sum $h + k$. Hence there is an addition table of the form

$$(7.1) \quad u(h) + u(k) = u(h + k) + f(h, k),$$

where $f(h, k)$ lies in G for each pair of elements h, k in H . The commutative and associative laws in the group E imply two corresponding identities for f ,

$$(7.2) \quad f(h, k) = f(k, h),$$

¹² Group extensions are discussed by Baer [2], Hall [6], Turing [11], Zassenhaus [15], and elsewhere. Much of the discussion in the literature treats the more general case in which G but not H is assumed to be abelian and in which G is not necessarily in the center of H .

$$(7.3) \quad f(h, k) + f(h + k, l) = f(h, k + l) + f(k, l).$$

The sum of any two elements $g_1 + u(h)$ and $g_2 + u(k)$ of E is determined by the addition table (7.1) and the addition given within G and H .

The extension E does not uniquely determine the corresponding function f . An arbitrary set of representatives $u'(h)$ for the elements of H can be expressed in terms of the given representatives as

$$u'(h) = u(h) + g(h), \quad \text{each } g(h) \in G;$$

they will have an addition table like that of (7.1) with a function f' given by

$$(7.4) \quad f'(h, k) = f(h, k) + [g(h) + g(k) - g(h + k)].$$

Conversely, a *factor set* of H in G is any function $f(h, k)$, with values in G for h, k in H which satisfies the "commutative" and "associative" conditions (7.2) and (7.3) for all h, k , and l in H . A *transformation set* is any function of h and k like the term in brackets in (7.4); thus for any function $g(h)$ defined for each $h \in H$ and taking on values in G , the function

$$(7.5) \quad t(h, k) = g(h) + g(k) - g(h + k)$$

is a transformation set. Such a set automatically satisfies the conditions (7.2) and (7.3), hence is always a factor set. Two factor sets f and f' are said to be *associate* if their difference is, as in (7.4), a transformation set. The correspondence between group extensions and factor sets may now be formulated as follows.

THEOREM 7.1. *For given groups G and H , there is a many-one correspondence $f \rightarrow (E, \beta)$ between the factor sets f of H in G and the group extensions (E, β) of G by H , where $f \rightarrow (E, \beta)$ holds if and only if f is the factor set which appears in one of the possible "addition tables" (7.1) for E . Two factor sets f and f' of H in G determine equivalent group extensions of G by H if and only if they are associate. In particular, the group extension determined by f is trivial if and only if f is a transformation set.*

PROOF. As a preliminary, observe that the associative relations (7.3) for f show (with $k = l = 0$, $h = k = 0$) that $f(0, 0) = f(h, 0) = f(0, l)$. Now, given f , we construct E_f as the group of all pairs (g, h) with addition given by the rule

$$(g_1, h) + (g_2, k) = (g_1 + g_2 + f(h, k), h + k),$$

and the homomorphism β_f defined by $\beta_f(g, h) = h$. Since $f(0, 0) = f(0, l)$, each element $(g, 0)$ may be identified with the corresponding element $g + f(0, 0)$ in G ; the pair (E_f, β_f) is then indeed an extension of G by H . As a representative of h in E_f , we may choose $u(h) = (0, h)$; the addition table (7.1) then involves exactly the original factor set f . If E is an arbitrary group extension

of G by H in which f appears as the factor set of E , the correspondence $g + u(h) \leftrightarrow (g, h)$ shows that the extension E is in fact equivalent to the extension E_f just constructed. Therefore $f \rightarrow (E_f, \beta_f)$ is a many-one correspondence with the defining property stated in the theorem.

If f and f' are associate, as in (7.4), the correspondence

$$(g, h) \rightarrow (g - g(h), h)'$$

shows that the corresponding extensions E_f and $E_{f'}$ are equivalent. Conversely, the argument leading to (7.4) shows in effect that E_f is equivalent to $E_{f'}$ only if f is associate to f' .

We turn now to two special applications of transformation sets. In the first place, the representative for the zero element of H may always be chosen as the zero in E . This means that $u'(0) = 0$, $u'(0) + u'(h) = u'(h)$, so that

$$(7.6) \quad f'(0, h) = f'(h, 0) = 0 \quad (\text{all } h \in H).$$

A factor set f' with the property (7.6) may be called *normalized*; we have proved that every factor set f is associate to a normalized factor set.

Free groups may be characterized in terms of group extensions as follows:

THEOREM 7.2. *A group with more than one element H is free if and only if every extension of any group by H is the trivial extension.*

PROOF. Suppose first that H satisfies the condition that every extension of every G is trivial. Represent H as F/R , where F is free. Then F is a trivial extension of R by H , hence is a direct sum of R and H . Therefore H , as a subgroup of the free group F , is itself free. The other half of the theorem is stated in more detail in the following Lemma.

LEMMA 7.3. *Every factor set f' of a free group F in a group G is a transformation set, so that*

$$(7.7) \quad f'(x, y) = \phi(x + y) - \phi(x) - \phi(y), \quad \phi(x) \in G,$$

holds for all $x, y \in F$. If F has generators z_α , the function ϕ may be chosen so that $\phi(0) = -f'(0, 0)$, $\phi(z_\alpha) = 0$ for each generator z_α .

PROOF. In the extension $E_{f'}$ of G by F we have an addition table

$$u'(x) + u'(y) = u'(x + y) + f'(x, y) \quad (x, y \in F).$$

In E we introduce a new set of representatives $u(\sum e_\alpha z_\alpha) = \sum e_\alpha u'(z_\alpha)$ for the elements $\sum e_\alpha z_\alpha$ of F . These are related to the original representatives by an equation $u(z) = u'(z) + \phi(z)$, where $\phi(z)$ has values in G . Because F is a free group, $z \rightarrow u(z)$ as defined is a homomorphism of F into E , so that $u(x + y) = u(x) + u(y)$, and the factor set belonging to u is identically zero. But the given f' is associate to this zero factor set, as in (7.4). Setting $f = 0$, $\phi = -g$ in (7.4) gives (7.7), as desired. By construction, $u(z_\alpha) = u'(z_\alpha)$, so $\phi(z_\alpha) = 0$. Also $u'(0) + u'(0) = u'(0) + f'(0, 0)$, so that $u'(0) = f'(0, 0)$, $u(0) = 0$, and therefore $\phi(0) = -f'(0, 0)$. This completes the proof.

8. The group of extensions

For fixed H and G the sum of two factor sets f_1 and f_2 is a third factor set, defined as

$$(f_1 + f_2)(h, k) = f_1(h, k) + f_2(h, k) \quad (h, k \in H).$$

Under this addition, the factor sets and the transformation sets form groups, denoted respectively by

$$(8.1) \quad \text{Fact } \{G, H\} = \text{group of all factor sets of } H \text{ in } G,$$

$$(8.2) \quad \text{Trans } \{G, H\} = \text{group of all transformation sets of } H \text{ in } G.$$

The factor sets belonging to a given group extension E constitute a coset of the subgroup $\text{Trans } \{G, H\}$, as in (7.4). Hence the correspondence of factor sets to extensions is a one-one correspondence between cosets of Fact/Trans and equivalence classes of extensions. This correspondence carries the addition of factor sets into an addition of group extensions. We are thus led to define the *group of group extensions* of G by H as¹³

$$(8.3) \quad \text{Ext } \{G, H\} = \text{Fact } \{G, H\} / \text{Trans } \{G, H\}.$$

If H is discrete while G is a (generalized) topological group, there will be a corresponding induced topology on $\text{Ext } \{G, H\}$. For each factor set f is a function on $H \times H$ with values in G , so that $\text{Fact } \{G, H\}$ is a subgroup of the generalized topological group $G^{H \times H}$ of all such functions. The subgroup "Trans" and the factor group "Ext" also carry topologies. Much as in §3 one can prove that if H is discrete and G topological, then $\text{Fact } \{G, H\}$ is a closed subgroup of $G^{H \times H}$. This proves

LEMMA 8.1. *If H is discrete and G is a topological (and compact) group, then $\text{Fact } \{G, H\}$ is a topological (and compact) group.*

In general, however, $\text{Trans } \{G, H\}$ will not be closed in $\text{Fact } \{G, H\}$, even when G is topological. In such cases $\text{Ext } \{G, H\}$ is necessarily a generalized topological group.

If (E, β) is an extension of G by H , each subgroup $S \subset H$ determines a corresponding subgroup $E_S \subset E$, consisting of all $e \in E$ with $\beta(e) \in S$. Since $E_S \supset G$, we may thus say that E "induces" an extension (E_S, β) of G by S . We call an extension E *finitely trivial* if E_S is trivial for every finite subgroup $S \subset H$.

Similarly, any factor set f of H in G determines for each subgroup $S \subset H$ a factor set f_S of S in G , where $f_S(h, k) = f(h, k)$ for h, k in S (i.e., f_S is obtained by "cutting off" f at S). The correspondence between factor sets and group extensions readily gives

LEMMA 8.2. *A factor set f of H in G determines a finitely trivial extension of*

¹³ It is possible to define the sum of two group extensions directly, without using the factor sets (see Baer [2] p. 394); it also is possible to give an analogous definition of the topology introduced below in $\text{Ext } \{G, H\}$.

G by H if and only if, for every finite subgroup $S \subset H$, the factor set f_S "cut off" at S is a transformation set of S in G . Hence the finitely trivial extensions of G by H constitute a subgroup $\text{Ext}_f\{G, H\}$ of $\text{Ext}\{G, H\}$.

9. Group extensions and generators

A group extension can be described not only by factor sets, but also by certain homomorphisms related to the generators of the extending group H . For let (E, β) be a given extension of G by H , and $H = F/R$ a representation of H as a factor group of a free group F . Let F have the generators z_α , as in §4; the corresponding elements (or cosets) t_α of H will then be a set of generators of H . For each generator t_α choose a corresponding representative u_α in the given group extension E , so that $\beta u_\alpha = t_\alpha$. Then $\beta(\sum e_\alpha u_\alpha) = \sum e_\alpha t_\alpha$, so that any element $\sum e_\alpha t_\alpha \in H$ has a representative of the form $\sum e_\alpha u_\alpha$. This means that each element of E can be written in the form

$$x = g + \sum e_\alpha u_\alpha, \quad g \in G, \quad e_\alpha \text{ integers.}$$

From this representation one can at once determine how to add the elements of E . However, this representation is not in general unique, for $(\sum e_\alpha u_\alpha) \in G$ is equivalent to $\sum e_\alpha t_\alpha = 0$, which in turn is equivalent to $(\sum e_\alpha z_\alpha) \in R$. Thus to each $r = \sum e_\alpha z_\alpha$ in the group R of "relations" there is assigned an element $\theta(r) \in G$, defined as

$$\theta(r) = \theta(\sum e_\alpha z_\alpha) = \sum e_\alpha u_\alpha$$

These assignments $\theta(r)$ completely determine the extension E .

The function θ hereby defined¹⁴ is a homomorphism of R into G . Conversely every such homomorphism θ can be used to construct a corresponding group extension of G by $H = F/R$; it suffices to construct E by reducing the direct product $F \times G$ modulo the subgroup of all elements of the form $(r, \theta(r))$, for $r \in R$. There is thus a correspondence between homomorphisms of R into G and extensions of G by $H = F/R$.¹⁵

10. The connection between homomorphisms and factor sets

Given G and $H = F/R$, an extension E of G by H may be given either by a factor set or by a homomorphism of R into G . There must therefore be a relation between factor sets and homomorphisms of this type. We now propose to establish this relation directly, without using extensions explicitly. (Actually, the correspondence which we obtain is identical with that obtained by going from a homomorphism first to the corresponding group extension and then to its factor set.)

THEOREM 10.1. *If $H = F/R$ is a factor group of a free group F , while G is any other group, then*

¹⁴ Actually θ may be obtained by "cutting off" one of the homomorphisms ϕ as described in Lemma 4.3.

¹⁵ This correspondence has been stated by Baer ([2], p. 395) and used by Hall [6].

$$(10.1) \quad \text{Ext } \{G, H\} \cong \text{Hom } \{R, G\} / \text{Hom } \{F \mid R, G\}.$$

Under the correspondence which gives this isomorphism

$$(10.2) \quad \text{Ext}_f \{G, H\} \cong \text{Hom}_f \{R, G; F\} / \text{Hom } \{F \mid R, G\},$$

$$(10.3) \quad \text{Ext } \{G, H\} / \text{Ext}_f \{G, H\} \cong \text{Hom } \{R, G\} / \text{Hom}_f \{R, G; F\}.$$

If G is a generalized topological group while F and H are discrete, all these isomorphisms are bicontinuous.

PROOF. As a preliminary, observe that the representation $H = F/R$ means that the free group F is a group extension of R by H . In this extension choose a representative $u_0(h)$ in F for each $h \in H$. F is then described, as in (7.1), by an addition table

$$(10.4) \quad u_0(h) + u_0(k) = u_0(h + k) + f_0(h, k),$$

where f_0 is a factor set of H in R . This factor set will be fixed throughout the proof.

Since $\text{Ext } \{G, H\}$ is defined as Fact/Trans , the required isomorphism (10.1) could be established by a suitable correspondence of homomorphisms to factor sets. Let $\theta \in \text{Hom } \{R, G\}$ be given, and define f_θ by

$$(10.5) \quad f_\theta(h, k) = \theta[f_0(h, k)] \quad (h, k \in H).$$

The requisite commutative and associative laws (7.2) and (7.3) for f_θ follow from those for f_0 , and the correspondence $\theta \rightarrow f_\theta$ is a homomorphism of $\text{Hom } \{R, G\}$ into $\text{Fact } \{G, H\}$, and therefore into $\text{Ext } \{G, H\}$.

Suppose next that θ can be extended to a homomorphism θ^* of F into G . This homomorphism applied to (10.4) gives

$$\theta^*[f_0(h, k)] = \theta^*[u_0(h)] + \theta^*[u_0(k)] - \theta^*[u_0(h + k)].$$

If we set $g(h) = \theta^*[u_0(h)]$, the result asserts that $\theta^*f_0 = \theta f_0 = f_\theta$ is a transformation set.

Conversely, suppose that f_θ is a transformation set, so that $f_\theta(h, k) = g(h) + g(k) - g(h + k)$ for some function g . Now any element in F can be written, in only one way, in the form $r + u_0(h)$, with r in R , h in H . We define $\theta^*(r + u_0(h))$ as $\theta(r) + g(h)$. Clearly θ^* is an extension of θ ; a straightforward computation with (10.4) shows that θ^* is actually a homomorphism. In this case, then, θ is extendable to F .

We know now that the correspondence $\theta \rightarrow f_\theta$ is an isomorphism of $\text{Hom } \{R, G\} / \text{Hom } \{F \mid R, G\}$ into a subgroup of $\text{Ext } \{G, H\}$. It remains to prove that it is a homomorphism onto. At this juncture we use for the first time the assumption that F is a free group. Let E be a given extension of G by H , with a factor set f which we can assume is normalized, as in (7.6). Let β_0 be the given homomorphism of F on H . Use f to define a factor set f' of F in G by the equation

$$(10.6) \quad f'(x, y) = f(\beta_0 x, \beta_0 y), \quad x, y \in F.$$

Since F is free, f' is a transformation set, so we can find, as in Lemma 7.3, a function $\phi(z)$ on F to G with

$$(10.7) \quad \phi(x + y) = \phi(x) + \phi(y) + f'(x, y).$$

In particular, if x and y lie in R , $\beta_0 x = \beta_0 y = 0$, and $f'(x, y) = f(0, 0) = 0$, because f is normalized. Thus ϕ , restricted to R , is a homomorphism $\theta = \phi|_R$ of R into G . Furthermore, if ϕ is applied to the addition table (10.4) for F , the property (10.7) gives

$$\phi[u_0(h) + u_0(k)] = \phi[u_0(h + k)] + \phi[f_0(h, k)],$$

where a term $f'(u_0(h + k), f_0(h, k))$, which would have entered by (10.7), is zero because f is normalized, $f_0(h, k) \in R$, and $\beta_0 f_0(h, k) = 0$. Now compute $f(h, k)$ for h, k in H . By (10.6),

$$\begin{aligned} f(h, k) &= f'(u_0(h), u_0(k)) \\ &= \phi[u_0(h) + u_0(k)] - \phi[u_0(h)] - \phi[u_0(k)] \\ &= \phi[u_0(h + k)] - \phi[u_0(h)] - \phi[u_0(k)] + \phi[f_0(h, k)], \end{aligned}$$

in virtue of the equation displayed just above. This equation asserts that f is associate to the factor set $\phi f_0 = \theta f_0$. In other words, given the normalized factor set f , we have constructed a homomorphism θ for which f is essentially θf_0 . This completes the proof of (10.1).

It is desirable to find a more explicit expression for this dependence of θ on f . A simple induction applied to (10.7) will show that, for z_i in F ,

$$\phi\left(\sum_{i=1}^n z_i\right) = \sum_{i=1}^n \phi(z_i) + \sum_{k=1}^{n-1} f'\left(\sum_{i=1}^k z_i, z_{k+1}\right).$$

If z_i is one of the generators z_α of F , then $\phi(z_i) = 0$, by Lemma 7.2. If $z_i = -z_\alpha$ is the negative of a generator, then by (10.7)

$$\phi(0) = \phi(z_\alpha + (-z_\alpha)) = \phi(z_\alpha) + \phi(-z_\alpha) + f'(z_\alpha, -z_\alpha),$$

so that $\phi(-z_\alpha) = -f'(z_\alpha, -z_\alpha)$. Now any element of F can be written as a finite linear combinations of generators and hence as a sum $\sum x_i$, where each x_i is either a generator or the negative of a generator z_α , and where any given generator may appear several times in this sum. In particular, for any element $r = \sum x_i$ in the subgroup R , the previous formula for ϕ becomes a formula for $\theta = \phi|_R$,

$$(10.8) \quad \theta\left(\sum_{i=1}^n x_i\right) = -\sum' f(\beta_0 x_i, -\beta_0 x_i) + \sum_{k=1}^{n-1} f\left(\sum_{i=1}^k \beta_0 x_i, \beta_0 x_{k+1}\right),$$

where β_0 is the given homomorphism of F into H , and where the sum \sum' is taken over those elements x_i which are the negatives of generators. The

essential feature of this formula is the fact that it expresses $\theta(r)$ for $r \in R$ as a sum of a finite number of values of the given factor set f of H in G .

Now consider the continuity of the correspondence $\theta \rightarrow f_\theta$ used to establish (10.1). It suffices to establish the continuity at 0. If U is any open set, containing zero, in $\text{Hom } \{R, G\}/\text{Hom } \{F | R, G\}$, there will be an open set V containing 0 in G and a finite set of elements $r_1, \dots, r_s \in R$ such that U contains the cosets of all homomorphisms θ with $\theta(r_i) \in V$, $i = 1, \dots, s$.

For a given f , the expressions $\theta(r_i)$ of (10.8) for these elements r will involve but a finite number of elements of the factor set f . Because of the continuity of addition in G , we can construct an open set U' in $\text{Fact } \{G, H\}$ such that each $\theta(r_i)$ does in fact lie in the given V . This establishes the continuity of the correspondence $f \rightarrow \theta$. The continuity of the inverse correspondence is obtained by a similar argument on the definition (10.5) of this correspondence.

It remains only to consider the formulas (10.2) and (10.3) on finitely trivial extensions. Let θ and its correspondent f_θ be given, and let $F_0 \supset R$ be any subgroup of F for which F_0/R is finite. A previous argument, applied to F_0 instead of F , shows that θ can be extended to a homomorphism of F_0 into G if and only if f_θ , regarded as a factor set for F_0/R in G , is a transformation set. But the subgroup $\text{Hom}_f \{R, G; F\}$ by definition consists of all those θ which are extendable to every such F_0 , while Ext_f by Lemma 8.2 is obtained from those factor sets which are transformation sets on every such subgroup F_0 . $\text{Hom}_f \{R, G; F\}/\text{Hom } \{F/R, G\}$ is the subgroup corresponding to $\text{Ext}_f \{G, H\}$ under $\theta \rightarrow f_\theta$. This proves (10.2) and with it (10.3). The continuity of the isomorphisms in this case follows from the continuity of the isomorphism (10.1).

For subsequent purposes we observe that the correspondence $\theta \rightarrow f_\theta$ obtained in this proof is essentially independent of the choice of the fixed factor set f_0 for H in R . Specifically, if f_0 is replaced by an associate factor set f'_0 , f_θ will be replaced also by an associate factor set, so that the corresponding element of $\text{Ext } \{G, H\}$ is not altered.

11. Applications

The representation of $\text{Ext } \{G, H\}$ as $\text{Hom } \{R, G\}/\text{Hom } \{F | R, G\}$ gives an immediate proof of the invariance of the latter group, as stated in Theorem 4.2 of Chapter I. There are a number of other simple corollaries.

COROLLARY 11.1. *For a direct product $H \times H'$,*

$$(11.1) \quad \text{Ext } \{G, H \times H'\} \cong \text{Ext } \{G, H\} \times \text{Ext } \{G, H'\}.$$

If G is a generalized topological group, the isomorphism is bicontinuous.

PROOF. If $H = F/R$ and $H' = F'/R'$, we may write $H \times H' = (F \times F')/(R \times R')$, where $F \times F'$, like F and F' , is free. Each homomorphism of $R \times R'$ into G determines homomorphisms θ and θ' of the subgroups R and R' into G , and this correspondence yields a (bicontinuous) isomorphism

$$\text{Hom } \{R \times R', G\} \cong \text{Hom } \{R, G\} \times \text{Hom } \{R', G\}.$$

Furthermore, under the same correspondence

$$\text{Hom} \{(F \times F') \mid (R \times R'), G\} \cong \text{Hom} \{F \mid R, G\} \times \text{Hom} \{F' \mid R', G\}.$$

These two relations yield a corresponding isomorphism between the respective factor groups such as $\text{Hom} \{R, G\} / \text{Hom} \{F \mid R, G\}$. By the fundamental theorem, the latter isomorphism is the one asserted in (11.1).

This conclusion can also be established without using homomorphisms, by a direct argument like that of Lemma 7.2. (Choose new representatives in E for elements of $H \times H'$ by setting $u'(hh') = u(h)u(h')$). Another simple argument directly with the factor sets will give a companion "direct product" representation,

$$(11.2) \quad \text{Ext} \{G \times G', H\} \cong \text{Ext} \{G, H\} \times \text{Ext} \{G', H\};$$

this isomorphism is also bicontinuous.

COROLLARY 11.2. *If H is a cyclic group of order m , then*

$$(11.3) \quad \text{Ext} \{G, H\} \cong G/mG, \quad (mG = \text{all } mg, \text{ for } g \in G).$$

This isomorphism is also bicontinuous.

This is a well known result, which can be derived directly from our main theorem. The cyclic group H can be written as $H = F/R$, where F is an infinite cyclic group with generator z , R the subgroup generated by mz . Then any $\theta \in \text{Hom} \{R, G\}$ is uniquely determined by the image $\theta(mz) = h$ of the generator mz . This correspondence $\theta \rightarrow h \pmod{mG}$ gives the isomorphism (11.3).

A similar representation can be found for any finite abelian group H , simply by representing H as a direct product of cyclic groups of orders $m_i, i = 1, \dots, t$. By Corollary 11.1, $\text{Ext} \{G, H\}$ is then isomorphic to the direct product of the groups G/m_iG . A similar decomposition applies if the abelian group H has a finite number of generators. The result may be stated as follows.

COROLLARY 11.3. *If H has a finite number of generators, and T is the subgroup of all elements of finite order in H , then $\text{Ext} \{G, H\} \cong \text{Ext} \{G, T\}$, algebraically and topologically. The latter group is a direct product of groups of the form G/mG .*

Theorem 7.2 (extensions by a free group are trivial) has an analogue for infinitely divisible groups. Recall that G is *infinitely divisible* if for each $g \in G$ and each integer $m \neq 0$ the equation $mx = g$ has a solution $x \in G$.

COROLLARY 11.4. *A group G is infinitely divisible if and only if every extension of G by any group is the trivial extension.*

PROOF. If G is not infinitely divisible, some $G/mG \neq 0$, so that there will be a non-trivial extension of G by a cyclic group, as in Corollary 11.2. Conversely, suppose G is infinitely divisible. If $R \subset F$ are groups, a transfinite induction will show that every homomorphism of R into G can be extended to a homomorphism into G of the larger group F . Therefore the subgroup $\text{Hom} \{F \mid R, G\}$ exhausts the group $\text{Hom} \{R, G\}$, and $\text{Ext} \{G, F/R\} = 0$.

COROLLARY 11.5. *If T is the subgroup of all elements of finite order in H , then*

$$(11.4) \quad \text{Ext } \{G, H\} / \text{Ext}_f \{G, H\} \cong \text{Ext } \{G, T\} / \text{Ext}_f \{G, T\}.$$

This isomorphism is bicontinuous (if G is a generalized topological group).

PROOF. In the representation $H = F/R$, let F_T denote the set of all elements of F of finite order modulo R . The group T then has the representation $T = F_T/R$, while F_T , as a subgroup of a free group, is itself free. Now the group $\text{Hom}_f \{R, G; F\}$ by definition consists of all homomorphisms extendable to subgroups of F finite over R ; as these subgroups are all contained in F_T , the group Hom_f is identical with $\text{Hom}_f \{R, G; F_T\}$. If both factor groups in (11.4) are now represented by groups of homomorphisms, as in (10.3), the result is immediate.

Observe that when T has only elements of finite order, the group $\text{Ext}_f \{G, T\}$, though it consists of extensions E of G by T trivial on every finite subgroup of T , can contain non-trivial extensions. This is illustrated by the following example. Let p be a prime, and G a group with generators g, h_1, h_2, \dots and relations $p^i h_i = g$, for $i = 1, 2, \dots$. In this group G the intersection of all the subgroups $p^i G$ is the group generated by g alone. Let T be the group of all rational numbers of the form a/p^i , reduced modulo 1. Then all elements of T have finite order, and T may be written as $T = F/R$, where F is a free group with generators z_1, z_2, \dots , and R the free subgroup generated by $p z_1, p z_2 - z_1, p z_3 - z_2, \dots$. (The homomorphism $F \rightarrow T$ maps z_i into $1/p^i$.)

To prove $\text{Ext}_f \{G, T\} \neq 0$ it suffices to find a $\theta \in \text{Hom}_f \{R, G; F\}$ which is not in $\text{Hom} \{F/R, G\}$. Such a θ is determined by setting $\theta(p z_1) = g$, $\theta(p z_{i+1} - z_i) = 0$, $i = 1, 2, \dots$. The definition $\theta^*(z_{n-i}) = p^i h_n$ will provide an extension θ^* of θ to the finite subgroup of F generated by z_1, \dots, z_n . However, suppose that θ had an extension ϕ to F . Then $\phi(p z_{i+1}) = \phi(z_i)$, so that $\phi(z_1) = p^n \phi(z_{n+1})$ for every n . This means that $\phi(z_1)$ is in every subgroup $p^n G$, hence has the form eg for an integer e . But then $g = \theta(p z_1) = p \phi(z_1) = ep g$ gives a contradiction. Therefore $\text{Ext}_f \{G, T\} \neq 0$ in this case. However, if G has no elements of finite order, one can prove easily that $\text{Ext}_f \{G, T\} = 0$, using Lemma 5.1 (see §17 below).

For several types of topological groups G , §5 gives information on the topology of the various relevant subgroups of $\text{Hom} \{R, G\}$. By the main theorem, the conclusions of Lemmas 5.3, 5.4, and 5.5 can now be rewritten as conclusions about the topology of $\text{Ext} \{G, H\}$, as follows.

COROLLARY 11.6. *If H is discrete and G a generalized topological group, the closure of the zero element in the generalized topological group $\text{Ext} \{G, H\}$ contains $\text{Ext}_f \{G, H\}$. If, in addition, every subgroup mG is closed in G , for $m = 2, 3, \dots$, then $\text{Ext}_f \{G, H\}$ is closed in $\text{Ext} \{G, H\}$.*

In particular, if H has no elements of finite order, then every extension of G by H is trivial on (the non-existent) finite subgroups of H , consequently $\text{Ext}_f \{G, H\} = \text{Ext} \{G, H\}$ and the closure of 0 is the whole group $\text{Ext} \{G, H\}$. This means that $\text{Ext} \{G, H\}$ carries the "trivial" (generalized) topology in which the only open sets are the whole group and the empty set.

COROLLARY 11.7. *If H is discrete and G compact and topological, then $\text{Ext}_f \{G, H\} = 0$ and $\text{Ext} \{G, H\}$ is itself a compact topological group.*

This conclusion is obtained from Lemma 8.1 and from Lemma 5.4.

12. Natural homomorphisms

The basic homomorphism $\eta(\theta) = f_\theta$ mapping elements θ of $\text{Hom} \{R, G\}$ into factor sets f , as in Theorem 10.1, is a "natural" one. Specifically, this means that the application of η "commutes" with the application of any homomorphism T to the free group F and its subgroup R . To state this more precisely, we need to consider first the homomorphisms which T induces on the groups $\text{Hom} \{R, G\}$ and $\text{Ext} \{G, H\}$.

Let F' be a free group with subgroup R' , T a homomorphism $z' \rightarrow Tz'$ of F' into the free group F such that $T(R') \subset R$. T induces a homomorphism of $H' = F'/R'$ into $H = F/R$. This induced homomorphism will be written with the same letter T , so that $T(g + R') = Tg + R$, for any coset $g + R'$.

Now consider $\theta \in \text{Hom} \{R, G\}$. Clearly the product $\theta' = \theta T$ is an element of $\text{Hom} \{R', G\}$, and the correspondence $\theta \rightarrow \theta'$ is a homomorphism T_h^* of $\text{Hom} \{R, G\}$ into $\text{Hom} \{R', G\}$. Furthermore $\theta \in \text{Hom} \{F | R, G\}$ implies $\theta T \in \text{Hom} \{F' | R', G\}$, so that T_h^* also induces a homomorphism T_h^* ,

$$(12.1) \quad T_h^* : \text{Hom} \{R, G\} / \text{Hom} \{F | R, G\} \rightarrow \text{Hom} \{R', G\} / \text{Hom} \{F' | R', G\}.$$

Similarly, consider $f \in \text{Fact} \{G, H\}$. The function f' defined by

$$f'(h', k') = f(Th', Tk') \quad (h', k' \in H')$$

is a factor set of H' in G , and the correspondence $f \rightarrow f'$ is a homomorphism T_*^* of $\text{Fact} \{G, H\}$ into $\text{Fact} \{G, H'\}$. Furthermore, $f \in \text{Trans} \{G, H\}$ implies $f' \in \text{Trans} \{G, H'\}$, so that T_*^* also induces a homomorphism T_*^* for the corresponding factor groups $\text{Ext} = \text{Fact}/\text{Trans}$,

$$(12.2) \quad T_*^* : \text{Ext} \{G, H\} \rightarrow \text{Ext} \{G, H'\}.$$

By the (dual) homomorphisms induced on Ext or Hom by T we always mean these homomorphisms T_h^* and T_*^* .

THEOREM 12.1. *Let T be a homomorphism of F' into F with $T(R') \subset R$, where $F \supset R$ and $F' \supset R'$ are free groups, while η (or η') is the homomorphism of $\text{Hom} \{R, G\}$ onto $\text{Ext} \{G, F/R\}$ established in the proof of Theorem 10.1. Then*

$$(12.3) \quad \eta' T_h^* = T_*^* \eta,$$

where T_h^* , T_*^* are the appropriate homomorphisms induced by T on Hom and Ext , respectively.

PROOF. The figure involved is

$$\begin{array}{ccc} \text{Hom} \{R, G\} & \xrightarrow{\quad \eta \quad} & \text{Ext} \{G, F/R\} \\ \downarrow T_h^* & & \downarrow T_*^* \\ \text{Hom} \{R', G\} & \xrightarrow{\quad \eta' \quad} & \text{Ext} \{G, F'/R'\} \end{array}$$

The correspondence η was constructed from a factor set f_0 for F as an extension of R ; similarly, η' is based on a factor set f'_0 for H' in R' , such that

$$(12.4) \quad u'_0(h') + u'_0(k') = u'_0(h' + k') + f'_0(h', k'),$$

where $u'_0(h')$ is a representative of $h' \in H'$ in F' . First we determine the relation between f_0 and f'_0 . The given homomorphism T carries F' into F , H' into H and thus $u'_0(h')$ into $Tu'_0(h')$, a representative in F of Th' in H . This representative will differ from the given representative $u_0(Th')$ by an element of R , so that

$$Tu'_0(h') = u_0(Th') + \rho(h') \quad (\text{all } h' \text{ in } H').$$

where each $\rho(h')$ lies in R . Now the representatives $Tu'_0(h')$ will add with a factor set $Tf'_0(h', k')$, as may be seen by applying T to both sides of (12.4). This factor set in associate (in the group TH') to the originally given factor set f_0 of $H \supset TH'$; explicitly we have, by the argument leading to (7.4), that

$$Tf'_0(h', k') = f_0(Th', Tk') + [\rho(h') + \rho(k') - \rho(h' + k')].$$

Suppose now that $\theta \in \text{Hom } \{R, G\}$ is given. Application of η and then T_e^* will give, by the definitions of these correspondences, a factor set f' , with

$$\begin{aligned} f'(h', k') &= \theta[f_0(Th', Tk')] \\ &= \theta T[f'_0(h', k')] + [\theta\rho(h' + k') - \theta\rho(h') - \theta\rho(k')]. \end{aligned}$$

On the other hand, application of T_h^* and then η' will give, again by the appropriate definitions, a factor set f^* with

$$f^*(h', k') = \theta'[f'_0(h', k')] = \theta T[f'_0(h', k')].$$

Since $\theta\rho(h')$ is an element in G for each $h' \in H'$, these two equations show that f^* and f' are associate, hence that $f' = T_e^* \eta \theta$ and $f^* = \eta' T_h^* \theta$ do determine the same element of $\text{Ext } \{G, H\}$, as asserted in the theorem.

CHAPTER III. EXTENSIONS OF SPECIAL GROUPS

In this chapter we shall determine $\text{Ext } \{G, H\}$ more explicitly for various special groups G and H . We begin with a brief review of the theory of characters, which will be used extensively in this chapter and also in Chapters V and VI.

13. Characters¹⁶

Let G , H , and J be three generalized topological groups. G and H are said to be *paired* to J if a continuous function¹⁷ $\phi(g, h)$ with values in J is given

¹⁶ The character theory was discovered by Pontrjagin (see [8]), generalized by van Kampen (see Weil [12], Ch. VI and Lefschetz [7] Ch. II).

¹⁷ As a mapping $G \times H \rightarrow J$; for discussion of pairing, cf. [8], [14].

such that for any fixed g_0 , $\phi(g_0, h)$ is a homomorphism of H into J and for any fixed h_0 , $\phi(g, h_0)$ is a homomorphism of G into J .

Each subset $A \subset G$ determines a corresponding subset $\text{Annih } A \subset H$, called the *annihilator* of A , such that $h \in \text{Annih } A$ if and only if $\phi(g, h) = 0$ for all $g \in A$. Annihilators of subsets of H are defined similarly. It is clear that the annihilators are subgroups.

LEMMA 13.1. *If G and H are paired to a topological group J , then for each $A \subset G$, $\text{Annih } A$ is a closed subgroup of H .*

This is an immediate consequence of the continuity of ϕ for fixed g .

G and H are said to be *dually paired* to J if they are so paired that

$$\text{Annih } G = 0 \quad \text{and} \quad \text{Annih } H = 0.$$

LEMMA 13.2. *If G and H are paired to J then $G/\text{Annih } H$ and $H/\text{Annih } G$ are dually paired to J .*

The most important group pairings arise when $J = P$ is the additive group of reals reduced modulo 1. A homomorphism of a group G into P will be called a *character* and the group $\text{Hom } \{G, P\}$ will be written as $\text{Char } G$. Since P has no "arbitrarily small" subgroups, it follows from a remark in §3 that if G is compact, $\text{Char } G$ is discrete. Vice versa, by Corollary 3.2, if G is discrete, $\text{Char } G$ is compact and topological.

The basic lemma of the theory of characters is

LEMMA 13.3. *Let G be a discrete or compact topological group and let $g \neq 0$ be an element of G . There is then a character $\theta \in \text{Char } G$ such that $\theta(g) \neq 0$.*

In the case of discrete G the lemma follows easily from the proof of Corollary 11.4, since P is infinitely divisible. In the compact case the proof is much less elementary and uses the theory of invariant integration in compact groups.

The lemma can be equivalently formulated as follows:

LEMMA 13.4. *Let G be a discrete or compact topological group. G and $\text{Char } G$ are dually paired to P with the multiplication*

$$\phi(g, \theta) = \theta(g), \quad g \in G, \theta \in \text{Char } G.$$

Now let G and H be paired to P with $\phi(g, h)$ as multiplication. Since, for a fixed g , $\phi(g, h)$ is a character of H and, for fixed h , $\phi(g, h)$ is a character of G , we obtain induced mappings

$$(13.1) \quad G \rightarrow \text{Char } H, \quad H \rightarrow \text{Char } G.$$

A basic result of the character theory is

THEOREM 13.5. *Let the compact topological group G and the discrete group H be paired to P . The pairing is dual if and only if the induced mappings (13.1) are isomorphisms:*

$$G \cong \text{Char } H \quad \text{and} \quad H \cong \text{Char } G.$$

The following two theorems are consequences of the previous results:

THEOREM 13.6. *If G is a discrete or a compact topological group, then*

$$\text{Char Char } G \cong G.$$

THEOREM 13.7. *If the compact topological group G and the discrete group H are dually paired to P , then for every closed subgroup G_1 of G and every subgroup H_1 of H we have*

$$\text{Annih} [\text{Annih } G_1] = G_1, \quad \text{Annih} [\text{Annih } H_1] = H_1.$$

14. Modular traces

To study $\text{Ext } \{G, H\}$ for compact G we need a certain modification of the "trace" of an endomorphism of a free group. The simplest case of this modification refers to a correspondence which is not a homomorphism, but is a homomorphism, modulo m -folds of elements. It may be stated as follows.

LEMMA 14.1. *Let m be an integer, and let $r \rightarrow S(r)$ be a correspondence carrying the free group R into a finite subset of itself in such manner that*

$$(14.1) \quad S(r_1 + r_2) \equiv S(r_1) + S(r_2) \pmod{mR},$$

for all $r_1, r_2 \in R$. Let the elements y_α be any independent basis for R , and write $S(y_\alpha) = \sum_\beta c_{\alpha\beta} y_\beta$, with integral coefficients $c_{\alpha\beta}$. Then the "trace"

$$(14.2) \quad t_m(S) \equiv \sum_\alpha c_{\alpha\alpha} \pmod{m}$$

is a well defined finite integer, modulo m , independent of the choice of the basis y_α for R .

The proof is exactly parallel to the standard one (e.g. [1], p. 569) for an actual homomorphism of R to itself, using the "modular" homomorphism condition (14.1) at the appropriate junctures in place of the full homomorphism condition. A similar analogue of a special case of the "additivity" of traces will give the following conclusion.

LEMMA 14.2. *If in Lemma 14.1 the elements w_1, \dots, w_t are any independent elements of R such that $S(R)$ lies in the group generated by w_1, \dots, w_t , and if $S(w_i) = \sum_j d_{ij} w_j$, then $t_m(S) \equiv \sum_i d_{ii} \pmod{m}$.*

Now let R be a subgroup of the free group F , σ a homomorphism of R into a finite subgroup of F/R . There will then be at least one integer m for which $m\sigma(R) = 0$. Choose for each coset u of F/R a representative $\rho(u)$ in F ; then $\rho(u + v) \equiv \rho(u) + \rho(v) \pmod{R}$. For each $r \in R$, $m(\rho\sigma r)$ is also an element of R , and $S(r) = m(\rho\sigma r)$, where

$$R \xrightarrow{\sigma} F/R \xrightarrow{\rho} F \xrightarrow{m} R,$$

is a correspondence of R to R with the modular homomorphism property (14.1).¹⁸ The trace of the original homomorphism σ is now defined as

$$(14.3) \quad t(\sigma) \equiv t_m(S)/m \equiv t_m(m\rho\sigma)/m \pmod{1}.$$

¹⁸ S could also be described in terms of m and σ as follows: S is the essentially unique correspondence of R to a finite subset of $mF \cap R$ with property (14.1) and such that each $\sigma(r)$ is the coset modulo R of $S(r)/m$.

THEOREM 14.3. *If $R \subset F$, F a free group, and if σ is any homomorphism of R into a finite subgroup of F/R , then the trace $t(\sigma)$ defined by (14.3) is a unique real number, modulo 1, independent of the choices of m and ρ made in its definition. If σ_1 and σ_2 are two such homomorphisms of R to F/R ,*

$$(14.4) \quad t(\sigma_1 + \sigma_2) \equiv t(\sigma_1) + t(\sigma_2) \pmod{1}.$$

In particular, $t_0 \equiv 0 \pmod{1}$. Furthermore, if T_0 is a fixed finite subgroup of F/R , the correspondence $\sigma \rightarrow t(\sigma)$ is a continuous homomorphism of $\text{Hom } \{R, T_0\}$ into the reals modulo 1.

We are to prove the invariance of the definition of t . First, hold ρ fixed and replace m by a proper multiple $m' = km$. Then S and $t_m(S)$ are both multiplied by k , hence $t'(\sigma) \equiv t_{km}(kS)/km \equiv kt_m(S)/km \equiv t(\sigma)$ is unaltered, mod 1. Now hold m fixed and let ρ' be any second set of representatives $\rho'(u)$ for cosets $u \in F/R$. Then $\rho'(u) \equiv \rho(u) \pmod{R}$, so $S'(r) \equiv S(r) \pmod{mR}$, which implies that $t_m(S') \equiv t_m(S) \pmod{m}$. This shows that the trace is independent of ρ and m .

The additive property (14.4) is readily established; it is only necessary to choose a single integer in such a way that both $m\sigma_1 R$ and $m\sigma_2 R$ are zero.

Before establishing the continuity of $t(\sigma)$, we propose a more explicit representation of the finiteness of $t(\sigma)$. Let T_0 be a fixed finite subgroup of F/R , and choose a direct summand F_0 of F with a finite number of generators such that $F_0/(F_0 \cap R)$ contains T_0 . We can choose simultaneously ([1], p. 566) a basis z_1, \dots, z_n for F_0 and a basis y_1, \dots, y_s for $F_0 \cap R$ so that $y_i = d_i z_i$, for integers d_i , $i = 1, \dots, s \leq n$. Furthermore, one can prove $F_0 \cap R$ a direct summand of R ; there is then a (not necessarily denumerable) basis for R of the form $y_1, \dots, y_s, y_\alpha, y_\beta, \dots$. In particular, if $\sigma(R) \subset T_0$, we may choose $\rho(0) = 0$, $\rho(T_0) \subset F_0$, hence $S(R) = m\rho\sigma(R) \subset F_0 \cap R$. The equations for S and its trace then take the form

$$(14.5) \quad S(y_\gamma) = \sum_{i=1}^n c_{\gamma i} y_i, \quad t_m(S) \equiv \sum_{i=1}^n c_{ii} \pmod{m},$$

where $\gamma = 1, 2, \dots, s, \alpha, \beta, \dots$.

To prove $t(\sigma)$ continuous it suffices to establish the continuity at $\sigma = 0$, and hence to prove that $t(\sigma) \equiv 0$ for σ in a suitable neighborhood U of 0 in $\text{Hom } \{R, T_0\}$. Let U be the open set in $\text{Hom } \{R, T_0\}$ consisting of all σ with $\sigma(y_1) = \dots = \sigma(y_s) = 0$, where y_i is the special basis constructed from F_0 above. Then, because $\rho(0) = 0$, we have $S(y_i) = 0$, $t_m(S) \equiv 0 \pmod{m}$, and therefore $t(\sigma) \equiv 0 \pmod{1}$ for σ in U .

We next consider circumstances under which the traces will vanish.

LEMMA 14.4. *If $\sigma \in \text{Hom } \{R, F/R\}$ has an extension σ^* which carries F homomorphically into a finite subgroup T_0 of F/R , then $t(\sigma) \equiv 0 \pmod{1}$.*

PROOF. For T_0 we choose $y_i = d_i z_i$ as above, and then select ρ with $\rho(T_0) \subset F_0$ and m with $mT_0 = 0$ and each $d_i \equiv 0 \pmod{m}$. Then, for suitable integers e_{ij} ,

$$\rho\sigma^*(z_i) = \sum_{j=1}^n e_{ij} z_j, \quad i = 1, \dots, n;$$

furthermore $\rho\sigma^*(kz_i) \equiv k\rho\sigma^*(z_i) \pmod{R_0}$, for any integer k . But $S(y_i) = m\rho\sigma(y_i) = m\rho\sigma^*(d_i z_i) \equiv m d_i \rho\sigma^*(z_i) \pmod{mR_0}$. Then computing $t_m(S)$ by (14.5) and using the fact that $m \equiv 0 \pmod{d_j}$ for each j , we find that $t_m(S) \equiv m \sum e_{ii} \equiv 0 \pmod{m}$, as asserted.

Conversely, we can find certain circumstances in which the trace will assuredly not vanish.

LEMMA 14.5. *If $z \in F$ has order n , modulo R , and if σ is a homomorphism of R into the subgroup of F/R generated by the coset of z , then $\sigma(nz) \not\equiv 0$ implies $t(\sigma) \not\equiv 0 \pmod{1}$.*

PROOF. Let u denote the coset of z , modulo R . Choose the system of representatives so that $\rho(iu) = iz$, for $i = 0, \dots, n-1$, and use n as the integer m in the definition of the trace. Then $S = m\rho\sigma$ carries R into the cyclic subgroup generated by mz . Since $\sigma(nz) = ku$, where $k \not\equiv 0 \pmod{m}$, $S(nz) \equiv knz$, and the trace, as computed by Lemma 14.2, is $t_m(S) \equiv k \not\equiv 0 \pmod{m}$, as asserted.

15. Extensions of compact groups

The group of extensions of a compact topological group G can be expressed as an appropriate character group.

THEOREM 15.1. *If G is compact and topological, H discrete, then $\text{Ext}_f \{G, H\} = 0$ and there is a (bicontinuous) isomorphism:*

$$(15.1) \quad \text{Ext} \{G, H\} \cong \text{Char Hom} \{G, H\}.$$

If G_0 is the component of 0 in G and T the subgroup of all elements of finite order in H , then also

$$\text{Ext} \{G, H\} \cong \text{Char Hom} \{G, T\} \cong \text{Char Hom} \{G/G_0, T\}.$$

The last conclusion follows at once from the first, for $\text{Hom} \{G, H\}$ includes only continuous homomorphisms ϕ of the compact group G ; every such homomorphism must map the connected subgroup G_0 into 0. Furthermore each ϕ carries G into a finite subgroup of the discrete group H , hence into a subgroup of T . Observe also that H is discrete, hence has no arbitrarily small subgroups; therefore (cf. §3) $\text{Hom} \{G, H\}$ is discrete, as should be the case for a character group of the compact group $\text{Ext} \{G, H\}$.

It remains to prove (15.1). Represent H as F/R ; then, according to the fundamental theorem of Chapter II, (15.1) is equivalent to

$$(15.2) \quad \text{Hom} \{R, G\} / \text{Hom} \{F \mid R, G\} \cong \text{Char Hom} \{G, F/R\}.$$

According to Theorem 13.5 it will thus suffice to provide a suitable pairing of the compact group $\text{Hom} \{R, G\}$ and the discrete group $\text{Hom} \{G, F/R\}$ to the reals modulo 1. To this end, take any $\theta \in \text{Hom} \{R, G\}$ and $\phi \in \text{Hom} \{G, F/R\}$. As just above, $\phi(G)$ is a finite subgroup of F/R . Therefore $\sigma = \phi\theta$ is a homomorphism of R into a finite subgroup of F/R , so that the trace introduced in the previous section can be used to define

$$(15.3) \quad t(\theta, \phi) \equiv t(\phi\theta) \pmod{1}.$$

We propose to show that this is the requisite pairing.

In the first place, this product is additive, for

$$t(\theta + \theta', \phi) \equiv t(\theta, \phi) + t(\theta', \phi) \pmod{1},$$

$$t(\theta, \phi + \phi') \equiv t(\theta, \phi) + t(\theta, \phi') \pmod{1}$$

follow from the corresponding property (14.4) for $\sigma = \phi\theta$. Secondly, if ϕ is fixed, the product $t(\theta, \phi)$ is continuous in θ . For when ϕ is fixed, $\sigma = \phi\theta$ maps R into a fixed finite subgroup of F/R . Since $\theta \rightarrow \phi\theta = \sigma$ is continuous, and since $\sigma \rightarrow t(\sigma)$ is continuous, by Theorem 14.3, the continuity of $t(\theta, \phi)$ follows.

As to the annihilators under this pairing, we assert that

$$(15.4) \quad \text{Annih Hom } \{G, F/R\} = \text{Hom } \{F \mid R, G\}.$$

For suppose first that $\theta \in \text{Hom } \{F \mid R, G\}$ and let θ^* be an extension of θ to F . Then $\sigma^* = \phi\theta^*$ is an extension of $\sigma = \phi\theta$ to F , and σ^* still carries F into (the same) finite subgroup of F/R . Therefore, by Lemma 14.4, $t(\theta, \phi) \equiv t(\sigma) \equiv 0 \pmod{1}$. Hence θ is in the annihilator in question.

Conversely, let θ be fixed, and suppose that $t(\theta, \phi) \equiv 0 \pmod{1}$ for every ϕ ; then $\theta \in \text{Hom } \{F \mid R, G\}$. Since G is compact and topological, it will suffice by Lemma 5.4 to prove that $\theta \in \text{Hom}_f \{R, G; F\}$. If this were not the case, there would be in F an element z of some order n , modulo R , such that $\theta(nz) = g_0$ is not an element of nG . But nG is a continuous image (under $g \rightarrow ng$) of the compact group G , hence (Lemma 1.1) is a closed subgroup of G ; therefore G/nG is compact and topological. By Lemma 13.3 there is then character χ of G/nG with $\chi(g'_0) \neq 0$, where g'_0 is the coset of g_0 modulo nG . Since every coset of G/nG has as order some divisor of n , this character χ carries G/nG into the group generated by the fraction $1/n$, modulo 1. This is a cyclic group of order n , and so can be replaced by the isomorphic cyclic group of order n generated by the coset z' of z in F/R . The so-modified character X of G/nG then induces a continuous homomorphism ϕ of G into F/R , where

$$\phi(g_0) \neq 0, \quad \phi(G) \subset [0, z', z'^2, \dots, z'^{n-1}].$$

For this particular ϕ , the homomorphism $\sigma = \phi\theta$ carries nz into $\phi\theta(nz) = \phi(g_0) \neq 0$. Lemma 14.5 of the previous section then shows that $t(\sigma) \equiv t(\theta, \phi) \not\equiv 0 \pmod{1}$, contrary to the assumption $t(\theta, \phi) \equiv 0$ for every ϕ . Therefore θ does lie in $\text{Hom } \{F \mid R, G\}$, and 15.4 is proved.

Finally, we assert that, under the pairing t ,

$$(15.5) \quad \text{Annih Hom } \{R, G\} = 0.$$

For suppose instead that $t(\theta, \phi) \equiv 0 \pmod{1}$ for all θ and for some $\phi \neq 0$. Then for some $g_0 \in G$, $\phi(g_0) = u \neq 0$. The element u of F/R is the coset of some element w of F ; as before, ϕ maps G into a finite subgroup of F/R , so that w

has a finite order m , modulo R . It is then possible to select in the free group F an independent basis with a first element z_0 such that $w = kz_0$ for some integer k . If z_0 has order n , modulo R , there is then a corresponding basis for R of elements y_α , with $y_0 = nz_0$. Now construct $\theta \in \text{Hom}\{R, G\}$ by setting

$$\theta(y_0) = g_0, \quad \theta(y_\alpha) = 0, \quad y_\alpha \neq y_0.$$

This particular homomorphism carries R into the subgroup of G generated by g_0 , so that the product $\sigma = \phi\theta$ carries R into the finite subgroup of F/R generated by $\phi(g_0) = u$. Since u is the coset of $w = kz_0$, this is contained in the subgroup of F/R generated by the coset of z_0 . Furthermore $\sigma(nz_0) = u \neq 0$. Lemma 14.5 again applies to show that $t(\sigma) = t(\theta, \phi) \not\equiv 0 \pmod{1}$, counter to assumption.

Given the assertions (15.4) and (15.5) as to annihilators, it follows from Lemma 13.2 that the groups $\text{Hom}\{R, G\}/\text{Hom}\{F \mid R, G\}$ and $\text{Hom}\{G, F/R\}$ are dually paired. Formula (15.2) is then a consequence of Theorem 13.5.

16. Two lemmas on homomorphisms

A generalized topological group G is said to have no arbitrarily small subgroups if there is in G an open set V containing 0 but containing no subgroups other than the group consisting of 0 alone.

LEMMA 16.1. *If the discrete group T has no elements of infinite order and the generalized topological group G has no arbitrarily small subgroups, while G_0 is the same group with the discrete topology, then $\text{Hom}\{T, G\}$ and $\text{Hom}\{T, G_0\}$ have the same topology.*

PROOF. $\text{Hom}\{T, G\}$ and $\text{Hom}\{T, G_0\}$ are algebraically identical. The hypotheses on T insure that every finite set of elements of T generates a finite subgroup of T . A complete set of neighborhoods U of 0 in $\text{Hom}\{T, G\}$ may therefore be found thus: take a finite subgroup $T_0 \subset T$ and an open set V_0 in G containing 0, and let U consist of all homomorphisms θ with $\theta(T_0) \subset V_0$. In particular, if V_0 is contained in the special open set V of G which contains no proper subgroups, the subgroup $\theta(T_0)$ is zero, so that U consists of all θ with $\theta(T_0) = 0$. The special sets U so described also form a complete set of neighborhoods of 0 in $\text{Hom}\{T, G_0\}$. Therefore the two topologies on the group are equivalent.

LEMMA 16.2. *Let $F \supset R$ be a free (discrete) group, $G' \supset G$ a discrete group, while $\text{Hom}\{F, G'; R, G\}$ denotes the set of all homomorphisms $\phi \in \text{Hom}\{F, G'\}$ with $\phi(R) \subset G$. Then*

$$(16.1) \quad \text{Hom}\{F, G'; R, G\}/\text{Hom}\{F, G\} \cong \text{Hom}\{F/R, G'/G\}.$$

PROOF. Any homomorphism of F/R into G'/G may be regarded as a homomorphism of F into G'/G which carries R into zero (Lemma 3.3), so that (16.1) becomes

$$(16.2) \quad \text{Hom}\{F, G'; R, G\}/\text{Hom}\{F, G\} \cong \text{Hom}\{F, G'/G; R, 0\}.$$

For each $\phi \in \text{Hom } \{F, G'\}$ let ϕ^* be the corresponding homomorphism reduced modulo G , so that for $x \in F$, $\phi^*(x)$ is the coset of $\phi(x)$, modulo G . The correspondence $\phi \rightarrow \phi^*$ is a homomorphism mapping $\text{Hom } \{F, G'; R, G\}$ into $\text{Hom } \{F, G'/G; R, 0\}$. Furthermore $\phi^* = 0$ if and only if $\phi(F) \subset G$, or $\phi \in \text{Hom } \{F, G\}$. Therefore $\phi \rightarrow \phi^*$ provides an (algebraic) isomorphism of the left hand group in (16.2) to a subgroup of the right hand group.

Conversely, select a fixed basis z_α for the free group F , and for each coset $b \in G'/G$ pick a fixed representative element $\rho(b)$ in G' . For given $\sigma \in \text{Hom } \{F, G'/G; R, 0\}$, define a corresponding homomorphism $\phi = \phi(\sigma)$, for any $x = \sum k_\alpha z_\alpha \in F$, as

$$\phi\left(\sum_\alpha k_\alpha z_\alpha\right) = \sum_\alpha k_\alpha \rho[\sigma z_\alpha].$$

This is a homomorphism of F into G' . By construction, $\rho[\sigma z_\alpha]$ modulo G is just σz_α , hence $\phi(x)$, modulo G , is $\sigma(x)$, or $\phi^* = \sigma$. This implies that $\phi(R) \subset G$, and so that each σ is the correspondent of some ϕ in the homomorphism $\phi \rightarrow \phi^*$.

To show (16.2) bicontinuous, we first analyze the topology in the groups involved. By the definition of the topology in a factor group, we have to consider only open sets in $\text{Hom } \{F, G'; R, G\}$ which are unions of cosets of $\text{Hom } \{F, G\}$. If z_1, \dots, z_n is any finite selection from the fixed set of generators for F , the set $U(z_1, \dots, z_n)$ consisting of all ϕ with $\phi(z_1) \equiv \dots \equiv \phi(z_n) \equiv 0 \pmod{G}$ is such an open set, and contains $\phi_0 = 0$. We assert that any open set V containing 0 which is a union of cosets contains one of these sets U . For, given V , there will be elements x_1, \dots, x_m in F such that V contains all ϕ with $\phi x_i = 0$. Select generators z_1, \dots, z_n such that each x_i can be expressed in terms of z_1, \dots, z_n ; then V contains all ϕ with $\phi z_i = 0$. Moreover, if $\phi z_i \equiv 0 \pmod{G}$, there is a homomorphism ϕ_1 of F into G with $\phi z_i = \phi_1 z_i$; since $\phi - \phi_1 \in V$, since V is a union of cosets of $\text{Hom } \{F, G\}$, and since $\phi_1 \in \text{Hom } \{F, G\}$, we conclude that $\phi \in V$. Thus $V \supset U(z_1, \dots, z_n)$.

A similar but simpler argument for $\text{Hom } \{F, G'/G; R, 0\}$ will show that every open set containing zero in this group contains all σ with $\sigma z_1 = \dots = \sigma z_n = 0$, for a suitable set of the generators of F . The mapping $\sigma \rightarrow \phi$ carries open sets of this special type into the open sets $U(z_1, \dots, z_n)$ described above, and conversely. This shows that the correspondence $\phi \rightarrow \phi^*$ is continuous at 0, and hence everywhere.

17. Extensions of integers

Next we consider the case in which every element of H has finite order; we then write T instead of H for this group. The group of extensions of the integers by such a group T can be written as a group of characters. In case T is finite, the result is a generalization of Corollary 11.2, for in this case $\text{Char } T \cong T$.

THEOREM 17.1. *If T has only elements of finite order, and if I is the (additive) group of integers,*

$$(17.1) \quad \text{Ext}_f \{I, T\} = 0,$$

$$(17.2) \quad \text{Ext } \{I, T\} \cong \text{Char } T.$$

The methods used to establish this result apply with equal force if I is replaced by any discrete group G which has no elements of finite order. The group $\text{Char } T$ of homomorphisms of T into the group of reals modulo 1 must then be replaced by a group of homomorphisms of T into another group suitably constructed from G . In fact, any G with no elements of finite order can be embedded in an essentially unique discrete group G_∞ with the following properties:¹⁹

- (i) G_∞ has no elements of finite order,
- (ii) G_∞/G has only elements of finite order,
- (iii) G_∞ is infinitely divisible.

For any $g \in G_\infty$ and any integer m there is then a unique $h = g/m$ in G_0 with $mh = g$. The (discrete) factor group G_∞/G is the analogue of the topological group P' of rationals modulo 1. Specifically, if $G = I$, $G_\infty = I_\infty$ is the group of rational numbers, and G_∞/G is the group P' , but with a discrete topology. Since T has only elements of finite order, $\text{Char } T$ is $\text{Hom } \{T, P'\}$. But P' clearly has no arbitrarily small subgroups, so that the latter group, by Lemma 16.1, is identical (algebraically and topologically) with $\text{Hom } \{T, I_\infty/I\}$. The exact generalization of Theorem 17.1 is thus

THEOREM 17.2. *If T has only elements of finite order, while G is discrete and has no elements of finite order, and G_∞ is defined as above,*

$$(17.3) \quad \text{Ext}_f \{G, T\} = 0,$$

$$(17.4) \quad \text{Ext } \{G, T\} \cong \text{Hom } \{T, G_\infty/G\}.$$

The isomorphism is bicontinuous if G and G_∞/G are both discrete.

PROOF. If T is represented in the form $T = F/R$, for F free, the conclusions of this theorem can be reformulated, according to the fundamental theorem of Chapter II, as

$$(17.3a) \quad \text{Hom}_f \{R, G; F\} = \text{Hom } \{F \mid R, G\},$$

$$(17.4a) \quad \text{Hom } \{R, G\} / \text{Hom } \{F \mid R, G\} \cong \text{Hom } \{F/R, G_\infty/G\}.$$

Observe first that any homomorphism $\theta \in \text{Hom } \{R, G\}$ can be extended in a unique way to a homomorphism $\theta^* \in \text{Hom } \{F, G_\infty\}$. For, since every element of $T = F/R$ has finite order, every $z \in F$ has a finite order modulo R . For each such z pick an integer m such that $mz \in R$, and define

$$(17.5) \quad \theta^*(z) = (1/m)\theta(mz), \quad z \in F, mz \in R.$$

This definition of θ^* is independent of the choice of m , and does yield a homomorphism of F into G_∞ . Clearly it is the only such homomorphism extending the given θ .

Suppose now that $\theta \in \text{Hom}_f \{R, G; F\}$. Each element $z \in F$ then generates a

¹⁹ G_∞ could also be described as a tensor product; see §18.

finite subgroup of F/R , so θ can be extended to a homomorphism mapping z and R into G . This extension of θ must agree with the unique extension θ^* . This shows that $\theta^*(z) \in G$ for each z , so that θ^* is in fact a homomorphism of F into $G \subset G_\infty$, and $\theta \in \text{Hom} \{F | R, G\}$. This proves (17.3a).

As in §16, let $\text{Hom} \{F, G_\infty; R, G\}$ denote the group of all homomorphisms $\phi \in \text{Hom} \{F, G_\infty\}$ with $\phi(R) \subset G$. This is a topological group, under the usual specification (§1) that any open set in $\text{Hom} \{F, G_\infty; R, G\}$ is the intersection of this group with an open set in the topological group $\text{Hom} \{F, G_\infty\}$.

The correspondence $\phi \rightarrow \phi | R$ provides a bicontinuous isomorphism

$$(17.6) \quad \text{Hom} \{F, G_\infty; R, G\} \cong \text{Hom} \{R, G\}.$$

For, by Lemma 3.4, $\phi \rightarrow \phi | R$ is a continuous homomorphism. It is an isomorphism because each $\theta \in \text{Hom} \{R, G\}$ has a unique extension $\theta^* = \phi \in \text{Hom} \{F, G_\infty; R, G\}$, by (17.5). This inverse correspondence is also continuous; for if U is the open set consisting of all ϕ with $\phi z_i = g_i$, for given $z_i \in F$ and $g_i \in G_\infty$, $i = 1, \dots, n$, there is an open set U_m in $\text{Hom} \{R, G\}$ consisting of all θ with $\theta(mz_i) = mg_i$, where m is chosen so that each $mz_i \in R$ and each $mg_i \in G$. The correspondence $\theta \rightarrow \theta^*$ of (17.5) carries U_m into U . This proves (17.6).

The correspondence $\phi \rightarrow \phi | R$ maps the subgroup $\text{Hom} \{F, G\}$ of $\text{Hom} \{F, G_\infty; R, G\}$ onto $\text{Hom} \{F | R, G\}$. Hence (17.6) also yields an isomorphism

$$\text{Hom} \{F, G_\infty; R, G\} / \text{Hom} \{F, G\} \cong \text{Hom} \{R, G\} / \text{Hom} \{F | R, G\}.$$

On the other hand, Lemma 16.2 provides an isomorphism

$$\text{Hom} \{F, G_\infty; R, G\} / \text{Hom} \{F, G\} \cong \text{Hom} \{F/R, G_\infty/G\}.$$

These two combine to give the required isomorphism (17.4a).

It should be remarked that the results of this section can also be obtained by arguments directly on factor sets, without the interposition of the fundamental theorem of Chapter II. Specifically, to prove Theorem 17.2, one could consider an extension E of G by T , determined by a factor set $f(s, t)$ for $s, t \in T$. If $t \in T$ has order m , let $\phi_{\pi}(t) \equiv (1/m) \sum_i f(it, t) \pmod{G}$, where $i = 0, 1, \dots, m-1$. In this fashion E determines a homomorphism $\phi_{\pi} \in \text{Hom} \{T, G_\infty/G\}$. Conversely, given such a homomorphism ϕ , one may select for each $\phi(t) \in G_\infty/G$ a representative element $\phi'(t) \in G_\infty$ and construct the corresponding factor set as $f(s, t) = \phi'(s) + \phi'(t) - \phi'(s+t)$. These correspondences will establish (17.4). The device of constructing ϕ_{π} by summation over the terms of the factor set is an application of the so-called "Japanese homomorphism," as commonly used for (multiplicative) factor sets.

18. Tensor products

Some of our formulas can be expressed more easily by means of the tensor products introduced by Whitney [13]. If A and B are given discrete abelian groups the *tensor product* $A \circ B$ is a set whose elements are finite formal sums

$\sum a_i b_i$ of formal products $a_i b_i$, with each $a_i \in A$, $b_i \in B$. Two such elements are added simply by combining the two formal sums into a single sum. Two such elements are equal if and only if the second can be obtained from the first by a finite number of replacements of the forms $(a + a')b \leftrightarrow ab + a'b$, $a(b + b') \leftrightarrow ab + ab'$. The tensor product $A \circ B$ so defined is a discrete abelian group, and the multiplication $a \cdot b$ is a pairing of A and B to $A \circ B$.

In the special case when $B = G$ is a group containing no elements of finite order, and $A = R_0$ is the additive group of rational numbers, any sum $\sum a_i b_i$ can, by the distributive law, be rewritten as a single term $(r/s)b$, where s is a common denominator for the rational numbers a_i . This representation is essentially unique. Therefore $R_0 \circ G$ is simply the group G_∞ used in §17 above, and G_∞/G is $(R_0 \circ G)/G$ (for details, cf. Whitney [13], pp. 507-508).

The tensor product can equivalently be defined in terms of characters, in the following fashion:

THEOREM 18.1. *If A and B are (discrete) abelian groups,*

$$(18.1) \quad A \circ B \cong \text{Char Hom } \{B, \text{Char } A\}.$$

PROOF. This conclusion can also be written in the form

$$(18.2) \quad \text{Char } (A \circ B) \cong \text{Hom } \{B, \text{Char } A\}.$$

Since the group of characters is the group of homomorphisms into the group P of reals modulo 1, this conclusion is a special case (with $C = P$) of the following

LEMMA 18.2. *If A and B are discrete abelian groups, C any generalized (topological) abelian group, then there is a bicontinuous isomorphism*

$$(18.3) \quad \text{Hom } \{A \circ B, C\} \cong \text{Hom } \{B, \text{Hom } (A, C)\}.$$

PROOF. Let $\theta \in \text{Hom } \{A \circ B, C\}$ be given. For each $b \in B$, let $\phi_b(a) = \theta(ab)$. Then $\phi_b \in \text{Hom } (A, C)$. Let $\omega_\theta(b) = \phi_b$. Then $\omega_\theta \in \text{Hom } \{B, \text{Hom } (A, C)\}$, and the correspondence $\theta \rightarrow \omega_\theta$ is a homomorphism of $\text{Hom } \{A \circ B, C\}$ into $\text{Hom } \{B, \text{Hom } (A, C)\}$. One verifies readily that it is an (algebraic) isomorphism ($\omega_\theta = 0$ only if $\theta = 0$). Furthermore, it is an isomorphism onto the whole group $\text{Hom } \{B, \text{Hom } (A, C)\}$. For let any ω in the latter group be given, with $\omega(b) = \phi'_b \in \text{Hom } (A, C)$ for each $b \in B$. Then define

$$\theta_\omega(\sum a_i b_i) = \sum \phi'_b(a_i), \quad a_i \in A, b_i \in B.$$

One verifies that θ_ω is uniquely defined, under the identifications $(a + a')b \rightarrow ab + a'b$, $a(b + b') \rightarrow ab + ab'$ used in the definition of $A \circ B$. Furthermore, $\theta_\omega \in \text{Hom } \{A \circ B, C\}$, and $\theta_\omega \rightarrow \omega$ in the previously given correspondence. Therefore $\theta \rightarrow \omega_\theta$, $\omega \rightarrow \theta_\omega$ does yield the indicated isomorphism (18.3). The continuity of the isomorphism in both directions is readily established from these explicit formulas and the appropriate definitions of open sets in the given topologies of the groups concerned.

CHAPTER IV. DIRECT AND INVERSE SYSTEMS

The Čech homology groups for a space are defined as limits of certain "direct" and "inverse" systems of homology groups for finite coverings of the space (Chap. VI). In view of our representation of homology groups in terms of groups of homomorphisms and groups of group extensions we are led to consider limits of groups of this sort. We shall show that the limit of a group of homomorphisms is itself a group of homomorphisms (§21) and that the corresponding proposition holds in certain special cases for groups of group extensions (§22). In the general case, however, we must introduce a new group to represent the limit of a group of group extensions. This group can also be introduced as a limit of tensor products (§25).

19. Direct systems of groups

A directed set J is a partially ordered set of elements $\alpha, \beta, \gamma, \dots$ such that for any two elements α and β there always exists an element γ with $\alpha < \gamma$, $\beta < \gamma$. For each index α in a directed set J let H_α be a (discrete) group, and for each pair $\alpha < \beta$, let $\phi_{\beta\alpha}$ be a homomorphism of H_α into H_β . If $\phi_{\gamma\alpha} = \phi_{\gamma\beta}\phi_{\beta\alpha}$ whenever $\alpha < \beta < \gamma$, the groups H_α are said to form a *direct system* with the projections $\phi_{\beta\alpha}$.²⁰

Any direct system determines a unique (discrete) limit group $H = \varinjlim H_\alpha$ as follows. Every element h_α of one of the groups H_α is regarded as an element h_α^* of the limit H , and two elements h_α^*, h_β^* are equal if and only if there is an index γ , $\alpha < \gamma$, $\beta < \gamma$, with $\phi_{\gamma\alpha}h_\alpha = \phi_{\gamma\beta}h_\beta$. Two elements h_α^* and h_β^* in H are added by finding some γ with $\alpha < \gamma$, $\beta < \gamma$; the sum is then the element $h_\gamma^* = (\phi_{\gamma\alpha}h_\alpha + \phi_{\gamma\beta}h_\beta)^*$. Under this addition and equality, the elements h_α^* form a group $H = \varinjlim H_\alpha$. Each of the given groups H_α has a homomorphism $\phi_\alpha(h_\alpha) = h_\alpha^*$ into the limit group, and $\phi_\beta\phi_{\beta\alpha} = \phi_\alpha$, for $\alpha < \beta$.

In case each given projection $\phi_{\beta\alpha}$ is an isomorphism (of H_α into H_β), the limit group can be regarded as a "union" of the given groups: each group H_α has an isomorphic replica $\phi_\alpha H_\alpha$ within H , and H is simply the union of these subgroups.

A subset J' of the set J of indices α is said to be *cofinal* in J if for each index α there is in J' an α' with $\alpha < \alpha'$. The limit $\varinjlim_{\alpha \in J'} H_\alpha$, taken over any such cofinal subset, is isomorphic to the original limit H .

20. Inverse systems of groups

For each index α in a directed set let A_α be a (generalized topological) group, and for each $\alpha < \beta$ let $\psi_{\alpha\beta}$ be a (continuous) homomorphism of A_β in A_α . If $\psi_{\alpha\beta}\psi_{\beta\gamma} = \psi_{\alpha\gamma}$ whenever $\alpha < \beta < \gamma$, the groups A_α are said to form an *inverse system* relative to the projections $\psi_{\alpha\beta}$. Each inverse system determines a limit group $A = \varprojlim A_\alpha$. An element of this limit group is a set $\{a_\alpha\}$ of elements $a_\alpha \in A_\alpha$ which "match" in the sense that $\psi_{\alpha\beta}a_\beta = a_\alpha$ for each $\alpha < \beta$. The sum

²⁰ Direct (and inverse) systems were discussed in Steenrod [9], Lefschetz [7], Chap. I and II, and in Weil [12], Ch. I.

of two such sets is $\{a_\alpha\} + \{b_\alpha\} = \{a_\alpha + b_\alpha\}$; since the ψ 's are homomorphisms, this sum is again an element of the group. This limit group A is a subgroup of the direct product of the groups A_α . The topology of the direct product $\prod A_\alpha$ thus induces (§1) a topology in $\varprojlim A_\alpha$; an open set in the latter group is the intersection with $\varprojlim A_\alpha$ of an open set of $\prod A_\alpha$. This makes $\varprojlim A_\alpha$ a generalized topological group. As before, a cofinal subset of the indices gives an isomorphic limit group.

Let each B_α be a subgroup of the corresponding group A_α of an inverse system, and assume, for $\alpha < \beta$, that $\psi_{\alpha\beta} B_\beta \subset B_\alpha$. Then the system B_α is an inverse system under the same projections $\psi_{\alpha\beta}$, and the limit $\varprojlim B_\alpha$ is, in natural fashion, a subgroup of $\varprojlim A_\alpha$. On the other hand, $\psi_{\alpha\beta}$ induces a homomorphism $\psi'_{\alpha\beta}$ of the (generalized topological) group A_β/B_β into A_α/B_α . Relative to these projections, the factor groups themselves form an inverse system A_α/B_α . The limit group of the latter system contains a homomorphic image of $\varprojlim A_\alpha$; if each a_α in A_α determines a coset a'_α in A_α/B_α , the map $\{a_\alpha\} \rightarrow \{a'_\alpha\}$ is a homomorphism of $\varprojlim A_\alpha$ into $\varprojlim (A_\alpha/B_\alpha)$ in which exactly the elements of $\varprojlim B_\alpha$ are mapped on zero. Thus we have

$$(20.1) \quad \varprojlim A_\alpha / \varprojlim B_\alpha \subset \varprojlim (A_\alpha / B_\alpha).$$

For compact topological subgroups this is an isomorphism:

LEMMA 20.1. *If the A_α form an inverse system relative to the $\psi_{\alpha\beta}$, and if each B_α is a compact topological subgroup of A_α with $\psi_{\alpha\beta} B_\beta \subset B_\alpha$, then*

$$(20.2) \quad \varprojlim A_\alpha / \varprojlim B_\alpha \cong \varprojlim (A_\alpha / B_\alpha).$$

PROOF. Consider any $c = \{c_\alpha\}$ in $\varprojlim (A_\alpha / B_\alpha)$, where $\psi'_{\alpha\beta} c_\beta = c_\alpha$ for each $\alpha < \beta$. Each $c_\alpha \in A_\alpha / B_\alpha$ is a coset of the compact topological subgroup B_α , hence itself is a compact Hausdorff subspace of the space A_α . Furthermore $\psi_{\alpha\beta}$ is a continuous mapping of the set c_β into c_α , for each $\alpha < \beta$. Since $\psi_{\alpha\gamma} = \psi_{\alpha\beta} \psi_{\beta\gamma}$, the sets c_α form an inverse system of compact non-empty Hausdorff spaces. Their limit space is therefore²¹ non-vacuous. This means that there is a set of elements $a_\alpha \in c_\alpha$ with $\psi_{\alpha\beta} a_\beta = a_\alpha$ for $\alpha < \beta$. The element $\{a_\alpha\}$ in the group $\varprojlim A_\alpha$ is therefore an element which maps onto the given element $\{c_\alpha\}$ in the homomorphism $\{a_\alpha\} \rightarrow \{a'_\alpha\}$ used to establish (20.1). The continuity of (20.2), in both directions, follows readily.

There is also an "isomorphism" theorem for inverse systems.

LEMMA 20.2. *If the groups A_α form an inverse system relative to the projections $\psi_{\alpha\beta}$, while C_α form an inverse system (with the same set of indices) relative to projections $\phi_{\alpha\beta}$, and if σ_α are (bicontinuous) isomorphisms of A_α to C_α , for every α , such that the "naturality" condition $\sigma_\alpha \psi_{\alpha\beta} = \phi_{\alpha\beta} \sigma_\beta$ holds, then the groups $\varprojlim A_\alpha$ and $\varprojlim C_\alpha$ are bicontinuously isomorphic.*

²¹ See Lefschetz [7], Theorem 39.1 or Steenrod [9], p. 666. Observe, however, that the latter proof is incomplete, because of the gap in lines 10-11 on p. 666.

21. Inverse systems of homomorphisms

Consider the group of all homomorphisms of H into G . As in Chap. II, §12, each projection $\phi_{\beta\alpha}$ of a direct system of groups H_α will induce a "dual" homomorphism $\phi_{\alpha\beta}^*$ of $\text{Hom } \{H_\beta, G\}$ into $\text{Hom } \{H_\alpha, G\}$. Furthermore $\phi_{\alpha\beta}^* \phi_{\beta\gamma}^* = \phi_{\alpha\gamma}^*$ for all $\alpha < \beta < \gamma$, so that the groups $\text{Hom } \{H_\alpha, G\}$ form an inverse system relative to these dual projections.

THEOREM 21.1. *If the (discrete) groups H_α form a direct system, then*

$$(21.1) \quad \text{Hom } \{\varprojlim H_\alpha, G\} \cong \varprojlim \text{Hom } \{H_\alpha, G\}.$$

PROOF. Consider any element $\omega = \{\theta_\alpha\}$ in $\varprojlim \text{Hom } \{H_\alpha, G\}$. To define a corresponding homomorphism θ_ω on $H = \varinjlim H_\alpha$, represent each element $h \in H$ as a projection $h = \phi_\alpha h_\alpha$ of some element $h_\alpha \in H_\alpha$, and set

$$(21.2) \quad \theta_\omega(h) = \theta_\omega(\phi_\alpha h_\alpha) = \theta_\alpha(h_\alpha), \quad h = \phi_\alpha h_\alpha.$$

The "matching" requirement that $\theta_\alpha = \phi_{\alpha\beta}^* \theta_\beta$ for $\alpha < \beta$ readily shows that $\theta_\omega(h)$ has a unique value, independent of the representation $h = \phi_\alpha h_\alpha$ chosen. Furthermore, $\theta_\omega \in \text{Hom } \{H, G\}$, and the correspondence $\omega \rightarrow \theta_\omega$ is an isomorphism.

Conversely, let any $\theta \in \text{Hom } \{H, G\}$ be given, and define

$$(21.3) \quad \theta_\alpha(h_\alpha) = \theta(\phi_\alpha h_\alpha), \quad h_\alpha \in H_\alpha.$$

If $\alpha < \beta$, $\phi_{\alpha\beta}^* \theta_\beta(h_\alpha) = \theta_\beta[\phi_{\beta\alpha} h_\alpha] = \theta[\phi_\beta \phi_{\beta\alpha} h_\alpha] = \theta(\phi_\alpha h_\alpha) = \theta_\alpha h_\alpha$; so $\phi_{\alpha\beta}^* \theta_\beta = \theta_\alpha$, and these θ 's match. Therefore $\omega = \{\theta_\alpha\}$ is an element of the inverse limit group $\varprojlim \text{Hom } \{H_\alpha, G\}$, and clearly θ_ω is the original homomorphism θ . The correspondence $\omega \rightarrow \theta_\omega$ therefore does establish the desired isomorphism (21.1). The continuity in both directions follows directly from the formulas (21.2) and (21.3) and the appropriate definition of neighborhoods of zero in the groups concerned.

22. Inverse systems of group extensions

Consider a direct system of discrete groups H_α . As in Chap. II, §12, each projection $\phi_{\beta\alpha}$ of H_α into H_β will induce a homomorphism $\phi_{\alpha\beta}^*$ of $\text{Ext } \{G, H_\beta\}$ into $\text{Ext } \{G, H_\alpha\}$. Furthermore $\phi_{\alpha\beta}^* \phi_{\beta\gamma}^* = \phi_{\alpha\gamma}^*$ for all $\alpha < \beta < \gamma$, so that the groups $\text{Ext } \{G, H_\alpha\}$ form an inverse system. Contrary to the situation in the previous section, the limit group $\varprojlim \text{Ext } \{G, H_\alpha\}$ may not be isomorphic to $\text{Ext } \{G, \varinjlim H_\alpha\}$. An example to this effect will be given below. However, there are two important cases when "Lim" and "Ext" are interchangeable.

THEOREM 22.1. *If G is compact and topological, while the (discrete) groups H_α form a direct system, then*

$$(22.1) \quad \text{Ext } \{G, \varinjlim H_\alpha\} \cong \varprojlim \text{Ext } \{G, H_\alpha\}.$$

This is proved by repeated applications of Lemma 20.1 to the representation

$$(22.2) \quad \text{Ext } \{G, H\} = \text{Fact } \{G, H\} / \text{Trans } \{G, H\},$$

where $H = \varinjlim H_\alpha$. Recall that any $f \in \text{Trans } \{G, H\}$ has the form

$$f(h, k) = g(h) + g(k) - g(h + k), \quad h, k \in H.$$

Here $g \in G^H$ is any mapping of H into G . Clearly $f = 0$ if and only if $g \in \text{Hom } \{H, G\}$, so

$$(22.3) \quad \text{Trans } \{G, H\} \cong G^H / \text{Hom } \{H, G\}.$$

The correspondence $g \rightarrow f$ is clearly continuous; since the isomorphism (22.3) is one-one and since the groups G^H and $\text{Hom } \{H, G\}$ are compact, by Lemma 3.1, the bicontinuity of (22.3) follows. Furthermore, this isomorphism is a "natural" one relative to homomorphisms, so that the isomorphism theorem for inverse systems (Lemma 20.2) gives

$$\varinjlim \text{Trans } \{G, H_\alpha\} \cong \varinjlim [G^{H_\alpha} / \text{Hom } \{H_\alpha, G\}].$$

In this representation the groups G^{H_α} and $\text{Hom } \{H_\alpha, G\}$ with the "dual" projections $\phi_{\alpha\beta}^*$ form inverse systems with the respective limits G^H and $\text{Hom } \{H, G\}$. Furthermore each group $\text{Hom } \{H_\alpha, G\}$ is compact and topological, so Lemma 20.1 gives

$$(22.4) \quad \begin{aligned} \varinjlim \text{Trans } \{G, H_\alpha\} &\cong \varinjlim G^{H_\alpha} / \varinjlim \text{Hom } \{H_\alpha, G\} \\ &= G^H / \text{Hom } \{H, G\} \cong \text{Trans } \{G, H\}. \end{aligned}$$

On the other hand one may show exactly as in the proof of Theorem 21.1 on homomorphisms that there is a bicontinuous isomorphism

$$(22.5) \quad \varinjlim \text{Fact } \{G, H_\alpha\} \cong \text{Fact } \{G, H\}.$$

Furthermore, each of the groups $\text{Trans } \{G, H_\alpha\}$ is compact and topological, so that Lemma 20.1 applies again to prove

$$\varinjlim [\text{Fact} / \text{Trans}] \cong \varinjlim \text{Fact} / \varinjlim \text{Trans}.$$

This, with (22.4) and (22.5), gives the desired conclusion.²²

THEOREM 22.2. *If G is discrete and has no elements of finite order, while T_α is a direct systems of discrete groups with only elements of finite order, then*

$$(22.6) \quad \text{Ext } \{G, \varinjlim T_\alpha\} \cong \varinjlim \text{Ext } \{G, T_\alpha\}.$$

The proof appeals directly to the result found in Theorem 17.2 of Chapter III, to the effect that

$$(22.7) \quad \text{Ext } \{G, T_\alpha\} \cong \text{Hom } \{T_\alpha, G_\infty / G\}.$$

The groups $\text{Hom } \{T_\alpha, G_\infty / G\}$ will form an inverse system under the dual projections $\phi_{\alpha\beta}^*$; as in Theorem 21.1 we then have

$$\text{Hom } \{\varinjlim T_\alpha, G_\infty / G\} \cong \varinjlim \text{Hom } \{T_\alpha, G_\infty / G\}.$$

²² Theorem 22.1 can also be proved by representing Ext by means of $\text{Char Hom } \{G, H\}$ as in Theorem 15.1. This argument, however, requires a tedious proof that the isomorphism established in the latter theorem is "natural," in the sense of §12.

But the group on the left is simply $\text{Ext } \{G, \varinjlim T_\alpha\}$, by another application of Theorem 17.2. The desired result should then follow by taking (inverse) limits on both sides in (22.7).

To carry out this argument, it is necessary to have the naturality condition which gives the isomorphism theorem (Lemma 20.2) for inverse systems. This naturality condition requires that the isomorphism (22.7) permute with the projections of the inverse systems. This is just a statement of the fact that the isomorphism (22.7) established in Theorem 17.2 is "natural" in the sense envisaged in §12. The proof of this naturality is straightforward, so details will be omitted.

COROLLARY 22.3. *If the discrete group G has only a finite number of generators, while T_α is a direct system of discrete groups with only elements of finite order, then*

$$\text{Ext } \{G, \varinjlim T_\alpha\} \cong \varinjlim \text{Ext } \{G, T_\alpha\}.$$

PROOF. Write G as $F \times L$ where F is free, L is finite (and thus compact). By (11.2) there is a "natural" isomorphism

$$\text{Ext } \{G, \varinjlim T_\alpha\} \cong \text{Ext } \{F, \varinjlim T_\alpha\} \times \text{Ext } \{L, \varinjlim T_\alpha\}.$$

The asserted result now follows by applying Theorem 22.2 to the first factor on the right, and Theorem 22.1 to the second, using Lemma 20.2.

We now show by an example that "Ext" and "Lim" do not necessarily commute. Let p be a fixed prime number, H the additive group of all rationals with denominator a power of p , and H_n the subgroup consisting of all multiples of $1/p^n$. Then $\varinjlim H_n = H$, since H is the union of the groups H_n . Furthermore H_n is a free group, so $\text{Ext } \{I, H_n\} = 0$, where I is the group of integers. On the other hand, $\text{Ext } \{I, \varinjlim H_n\} = \text{Ext } \{I, H\}$ is a group computed in appendix B; it is decidedly not zero, in fact it is not even denumerable.

23. Contracted extensions

Before further consideration of the inverse limits of groups of extensions, we make a comparison of the group of extensions of a group G by a group H with the group of extensions by a subgroup H_0 of H . The identity mapping I of H_0 into H is a homomorphism, hence, as in §12, will give dual homomorphisms

$$(23.1) \quad I^*: \text{Fact } \{G, H\} \rightarrow \text{Fact } \{G, H_0\},$$

$$(23.2) \quad I^*: \text{Trans } \{G, H\} \rightarrow \text{Trans } \{G, H_0\}.$$

Specifically, I^* is the operation of "cutting off" a factor set $f \in \text{Fact } \{G, H\}$ to give a factor set $f_0 = I^*f \in \text{Fact } \{G, H_0\}$; $f_0(h, k)$ is defined only for $h, k \in H_0$, and always equals $f(h, k)$. Clearly I^* carries transformation sets into transformation sets, as in (23.2). Thus I^* also induces a dual homomorphism

$$(23.3) \quad I^*: \text{Ext } \{G, H\} \rightarrow \text{Ext } \{G, H_0\}.$$

This homomorphism may be visualized as follows: given E such that $G \subset E$

and $E/G = H$, there is an $E_0 \subset E$ such that $G \subset E_0$ and $E_0/G = H_0$. Then $I^*(E) = E_0$.

LEMMA 23.1. *If H_0 is a subgroup of the group H then for any group G the homomorphism I^* of (23.3) maps the group $\text{Ext } \{G, H\}$ onto $\text{Ext } \{G, H_0\}$.*

PROOF.²³ Represent H as F/R , where F is free. There is then a subgroup F_0 of F such that $R \subset F_0$ and $F_0/R = H_0$. By the fundamental theorem we have isomorphisms

$$\text{Ext } \{G, H\} \cong \text{Hom } \{R, G\} / \text{Hom } \{F | R, G\},$$

$$\text{Ext } \{G, H_0\} \cong \text{Hom } \{R, G\} / \text{Hom } \{F_0 | R, G\},$$

where $\text{Hom } \{F | R, G\} \subset \text{Hom } \{F_0 | R, G\}$. According to the "naturality" theorem of §12 the homomorphism I^* between the groups on the left can be represented on the right as that correspondence which carries each coset of $\text{Hom } \{F | R, G\}$ into the coset of $\text{Hom } \{F_0 | R, G\}$ in which it is contained. This makes it obvious that the homomorphism is a mapping "onto."

LEMMA 23.2. *If $H_0 \subset H$, then the dual homomorphisms I^* of factor and transformation sets, as in (23.1) and (23.2), are mappings "onto."*

PROOF. Any element in $\text{Trans } \{G, H_0\}$ has the form

$$f(h, k) = g(h) + g(k) - g(h + k),$$

where g is an arbitrary function on H_0 to G . Let g^* be an arbitrary extension of g to H , and

$$f^*(h, k) = g^*(h) + g^*(k) - g^*(h + k).$$

Then f^* is a transformation set with $I^*f^* = f$. This proves that (23.2) is a mapping onto. Since (23.3) and (23.2) are mappings onto, the same holds for (23.1).

24. The group Ext^*

Since limits do not always permute with groups of extensions, we now introduce a new group which is the limit of an inverse system of groups of group extensions.

Consider a discrete group T with only elements of finite order. The set $\{S_\alpha\}$ of all finite subgroups of T is a direct system, if $\alpha < \beta$ means that $S_\alpha \subset S_\beta$, and that the projection $I_{\beta\alpha}$ of S_α into S_β is simply the identity. The direct limit of $\{S_\alpha\}$ is the group T .

Let G be any generalized topological group. Since $\{S_\alpha\}$ is a direct system, it follows from a previous section that the groups $\text{Ext } \{G, S_\alpha\}$ form an inverse system with projections $I_{\alpha\beta}^*$. We define our new group as the limit of this system

²³ The lemma can also be proved directly in terms of the group extensions E, E_0 , using a suitable transfinite induction.

$$(24.1) \quad \text{Ext}^* \{G, T\} = \varinjlim \text{Ext} \{G, S_\alpha\}.$$

The two theorems of §22 as to cases in which "Ext" and "Lim" commute give at once

COROLLARY 24.1. *If G is compact and topological, or is discrete without elements of finite order, then*

$$\text{Ext}^* \{G, T\} \cong \text{Ext} \{G, T\}.$$

In the definition of Ext^* we used the approximation of T by its finite subgroups S_α . However, any approximation by finite groups will give the same result:

THEOREM 24.2. *If T_α is any direct system of finite groups, the corresponding inverse system of groups $\text{Ext} \{G, T_\alpha\}$ has a limit*

$$(24.2) \quad \varinjlim \text{Ext} \{G, T_\alpha\} \cong \text{Ext}^* \{G, \varinjlim T_\alpha\}.$$

PROOF. In case T_α is the system of all finite subgroups of the limit $T = \varinjlim T_\alpha$, this equation is simply the definition of Ext^* . In general, $T = \varinjlim T_\alpha$ is a group in which every element has finite order. Each T_α has a homomorphic projection $T'_\alpha = \phi_\alpha T_\alpha$ into the limit T , and T is simply the union of these subgroups T'_α . The set of these subgroups T'_α is therefore cofinal in the set of all finite subgroups of T . The inverse system of the groups $\text{Ext} \{G, T'_\alpha\}$, relative to the "identity" projections $I_{\alpha\beta}^*$, is cofinal in the inverse system used to define Ext^* , hence gives the same limit group,

$$(24.3) \quad \text{Ext}^* \{G, T\} \cong \varinjlim \text{Ext} \{G, T'_\alpha\}.$$

An element f^* in this limit group can be represented (but not uniquely) as a set $\{f_\alpha\}$ of factor sets $f_\alpha \in \text{Fact} \{G, T'_\alpha\}$ which "match" modulo transformation sets. This means that for each $\beta > \alpha$ there is a transformation set $t_{\alpha\beta} \in \text{Trans} \{G, T'_\alpha\}$ such that

$$f_\alpha(h', k') = f_\beta(h', k') + t_{\alpha\beta}(h', k'), \quad h', k' \in T'_\alpha.$$

Now each homomorphism ϕ_α of T_α into T'_α determines, as in §12, a dual homomorphism ϕ_α^* of $\text{Fact} \{G, T'_\alpha\}$ into $\text{Fact} \{G, T_\alpha\}$, defined so that $e_\alpha = \phi_\alpha^* f_\alpha$ is the factor set given by the equations

$$(24.4) \quad e_\alpha(h, k) = f_\alpha(\phi_\alpha h, \phi_\alpha k), \quad h, k \in T_\alpha.$$

If the f_α match, one readily proves that the corresponding e_α also match, modulo transformation sets. If the representation of f^* by $\{f_\alpha\}$ is changed by adding to each f_α a transformation set, the e_α 's are changed accordingly by transformation sets. Therefore the correspondence

$$(24.5) \quad f^* = \{f_\alpha\} \rightarrow e^* = \{\phi_\alpha^* f_\alpha\} = \omega f^*$$

carries each element f^* in $\varinjlim \text{Ext} \{G, T'_\alpha\}$ into a well defined element e^* in $\varinjlim \text{Ext} \{G, T_\alpha\}$. One verifies at once that this correspondence is a homomorphism.

Now we use the assumption that each T_α is finite. If $\phi_\alpha h_\alpha = 0$ for some $h_\alpha \in T_\alpha$, the definition of equality in a direct system shows that $\phi_{\beta\alpha} h_\alpha = 0$ for some $\beta > \alpha$. Since the whole group T_α is finite, we can select a single $\beta = \beta_0(\alpha) > \alpha$ which will do this for all h_α , so that

$$\phi_\alpha h_\alpha = 0 \text{ implies } \phi_{\beta\alpha} h_\alpha = 0, \quad \beta = \beta_0(\alpha).$$

Since $\phi_\beta \phi_{\beta\alpha} = \phi_\alpha$, ϕ_β is now an isomorphism of $\phi_{\beta\alpha} T_\alpha$ onto T'_α . Let ϕ_β^{-1} denote the inverse correspondence.

Next we show that ω , as defined by (24.5), is an isomorphism. For suppose $\omega f^* = 0$; every $\phi_\alpha^* f_\alpha$ is then a transformation set t_α . Using (24.4) and $\beta = \beta_0(\alpha)$, we then have, for any $h', k' \in T'_\alpha$,

$$f_\alpha(h', k') \equiv f_\beta(h', k') = e_\beta(\phi_\beta^{-1} h', \phi_\beta^{-1} k') = t_\beta(\phi_\beta^{-1} h', \phi_\beta^{-1} k').$$

This shows that f_α is a transformation set, hence that $f^* = \{f_\alpha\} = 0$ in $\text{Ext}^* \{G, T\}$.

To construct a correspondence inverse to ω , let $e^* = \{e_\alpha\}$ be a given element in $\varinjlim \text{Ext} \{G, T_\alpha\}$, where each $e_\alpha \in \text{Fact} \{G, T_\alpha\}$. Define

$$(24.6) \quad f_\alpha(h', k') = e_\beta(\phi_\beta^{-1} h', \phi_\beta^{-1} k'), \quad \beta = \beta_0(\alpha)$$

for each $h', k' \in T'_\alpha$. Since the e_α 's are known to match, we may verify that the replacement of β by any larger index γ in this definition will only alter f_α by a transformation set. To show that f_α and f_γ match properly for $\alpha < \gamma$, one then chooses $\beta > \beta_0(\alpha)$, $\beta > \beta_0(\gamma)$ in (24.6) and uses the given matching of the e_α 's (modulo transformation sets). Finally, one verifies easily that the correspondence $\{e_\alpha\} \rightarrow \{f_\alpha\}$ of (24.6) is the inverse of the given correspondence ω of (24.5). This establishes the isomorphism (24.2) required in the theorem. The continuity, in both directions, follows from the formulae (24.5) and (24.6).

THEOREM 24.3. *If every element of T has finite order, the group $\text{Ext}^* \{G, T\}$ contains an everywhere dense subgroup isomorphic to $\text{Ext} \{G, T\} / \text{Ext}_f \{G, T\}$.*

This will be established by constructing a "natural" homomorphism of $\text{Ext} \{G, T\}$ into $\text{Ext}^* \{G, T\}$. To this end, let E be any extension of G by T determined by a factor set f . As in §23, f may be "cut off" to give a factor set f_α for any given finite subgroup $S_\alpha \subset T$. These factor sets match properly, so $\{f_\alpha\}$ determines a definite element in the inverse limit group $\text{Ext}^* \{G, T\}$. Alteration of f by a transformation set alters each f_α by the correspondingly "cut off" transformation set, hence does not alter the element $\{f_\alpha\} = f^*$ of Ext^* . Therefore $f \rightarrow \{f_\alpha\}$ is a well defined homomorphism of Ext into Ext^* . In case f lies in $\text{Ext}_f \{G, T\}$, each f_α is a transformation set, by the very definition of Ext_f , so that $\{f_\alpha\} = 0$. Conversely, if each f_α is a transformation set, $f \in \text{Ext}_f$. We thus have a (bicontinuous) isomorphism of $\text{Ext} / \text{Ext}_f$ onto a subgroup of Ext^* .

To show this subgroup everywhere dense in Ext^* it will suffice, whatever the topology in G , to show the following: Given an element $f^* = \{f'_\alpha\}$ in $\text{Ext}^* \{G, T\}$ and a finite set J_0 of indices, there exists a factor set f in $\text{Fact} \{G, T\}$ such that

$f_\alpha - f'_\alpha$ is a transformation set for every index $\alpha \in J_0$. To prove this, choose a finite subgroup S_γ which contains all the groups S_α , for $\alpha \in J_0$. By Lemma 23.2, the given factor set f'_γ can be obtained by "cutting off" a suitable factor set f , so that $f_\gamma - f'_\gamma$ is the transformation set 0. The matching condition for the f'_α then shows that each difference $f_\alpha - f'_\alpha$ is also a transformation set, for $\alpha \in J_0$. This proves the property stated above, and with it, the theorem.

In many cases the subgroup considered in Theorem 24.3 is the whole group Ext^* . It follows from previous considerations that this is the case when G is compact or when G is discrete and has no elements of finite order. Another important case is that when T is countable:

THEOREM 24.4. *If T is countable then*

$$(24.7) \quad \text{Ext}^* \{G, T\} \cong \text{Ext} \{G, T\} / \text{Ext}_f \{G, T\}.$$

PROOF. Since T is countable, the system of all finite subgroups of T used to define $\text{Ext}^* \{G, T\}$ may be replaced by a cofinal sequence of finite subgroups T_n with $T_1 \subset T_2 \subset \dots \subset T_n \subset \dots \subset T$, with the identity projections $I_n : T_n \rightarrow T_{n+1}$. Therefore $\text{Ext}^* \{G, T\} = \varprojlim \text{Ext} \{G, T_n\}$. An element e^* of this group can then be represented as a sequence $\{f_n\}$ of factor sets $f_n \in \text{Fact} \{G, T_n\}$ which match, in the sense that, for some g_n ,

$$(24.8) \quad f_{n+1}(h, k) = f_n(h, k) + [g_n(h) + g_n(k) - g_n(h + k)]$$

for all $h, k \in T_n$. The transformation set shown in brackets may be extended to all of T by extending g_n to a function g_n^* on T , as in Lemma 23.2. We introduce a new function $s_n(h) = g_1^*(h) + \dots + g_{n-1}^*(h)$, for all $h \in T$, and a new family of factor sets

$$f'_n(h, k) = f_n(h, k) - [s_n(h) + s_n(k) - s_n(h + k)],$$

for $h, k \in T_n$.²⁴ Since f'_n differs from f_n by a transformation set, the given element e^* of Ext^* has both representations $\{f_n\}$ and $\{f'_n\}$. But (24.8) also shows that f'_{n+1} , cut off at T_n , is exactly f'_n . Therefore these factor sets match exactly, and provide a composite factor set f of T in G . This factor set f is one which corresponds to the given element e^* of Ext^* in the "natural" homomorphism of Ext into Ext^* as constructed in Theorem 24.3, so this homomorphism maps Ext on all of Ext^* , as asserted in (24.7).

25. Relation to tensor products

The group Ext^* introduced in this chapter is closely related to the tensor product. Since an early form ([5]) of our results was formulated in terms of tensor products, we shall briefly state the connection. Let G be any group, A a compact zero-dimensional group, $\{A_\alpha\}$ the family of all open and closed subgroups of A . Then the groups A/A_α and $G \circ (A/A_\alpha)$ both form inverse

²⁴ This construction is an exact group theoretic analog of a similar matching process for chains, as devised by Steenrod ([9], p. 692).

systems. The modified tensor product $G \bullet A$ is defined as the limit of the groups $G \bullet (A/A_\alpha)$.

Now let the group T with all elements of finite order be represented in terms of a free group F as $T = F/R$. Each finite subgroup S_α then has a representation F_α/R , and the fundamental theorem of Chapter II asserts that

$$(25.1) \quad \text{Ext} \{G, S_\alpha\} \cong \text{Hom} \{R, G\} / \text{Hom} \{F_\alpha \mid R, G\}.$$

The groups on both sides here form inverse systems, relative to the identity as projections. Furthermore, the isomorphism of (25.1) permutes with these projections, so that the limits of the two direct systems in (25.1) are also isomorphic. In view of the definition of Ext^* , this gives

$$(25.2) \quad \text{Ext}^* \{G, T\} \cong \varprojlim [\text{Hom} \{R, G\} / \text{Hom} \{F_\alpha \mid R, G\}].$$

Now if I is the group of integers, any element $\sigma = \sum g_i \phi_i$ in the tensor product $G \bullet \text{Hom} \{R, I\}$ determines in natural fashion the homomorphism $\theta \in \text{Hom} \{R, G\}$ with $\theta(r) = \sum \phi_i(r) g_i$. By a somewhat lengthy argument, this correspondence $\sigma \rightarrow \theta$ can be used to "factor out" the G in (25.2) to give

$$(25.3) \quad \text{Ext}^* \{G, T\} \cong \varprojlim G \bullet [\text{Hom} \{R, I\} / \text{Hom} \{F_\alpha \mid R, I\}].$$

The group in brackets here is $\text{Ext} \{I, F_\alpha/R\}$, by the fundamental theorem on group extensions. According to Theorem 17.1 it can be expressed as $\text{Char } S_\alpha$. Therefore (25.3) is²⁵

$$(25.4) \quad \text{Ext}^* \{G, T\} \cong \varprojlim (G \bullet \text{Char } S_\alpha).$$

But the group $\text{Char } S_\alpha$ can, by the theory of characters (Lemma 13.2, Theorem 13.5), be rewritten as a factor group $\text{Char } T / \text{Annih } S_\alpha$, where the subgroups of the form $\text{Annih } S_\alpha$ in $\text{Char } T$ are exactly the open and closed subgroups in the zero-dimensional group $\text{Char } T$. Thus (25.4) may be restated in terms of the modified tensor product, as

$$(25.5) \quad \text{Ext}^* \{G, T\} \cong G \bullet \text{Char } T.$$

The use of the "modified" tensor product is therefore equivalent to the use of the group Ext^* .

CHAPTER V. ABSTRACT COMPLEXES

Turning now to the topological applications, we will establish the fundamental theorem on the decomposition of the homology groups of an infinite complex in terms of the integral cohomology groups of the complex. This theorem will be obtained in several closely related forms (Theorems 32.1, 32.2 and 34.2) for three different types of homology groups. The largest (or "longest") homology group is that consisting of infinite cycles, with coefficients in G , reduced modulo

²⁵ This argument requires an application of the isomorphism theorem for inverse systems, and hence rests on the fact that the isomorphism of Theorem 17.1 is "natural" in the sense of §12.

the subgroup of actual boundaries. Since the latter subgroup is not in general closed, this homology group will be only a generalized topological group. This suggests the introduction of the shorter "weak" homology group, which consists of cycles modulo "weak" boundaries; i.e. those cycles which can be regarded as boundaries in any finite portion of the complex. The fundamental theorem for this type of homology group uses the group Ext_f which has been already analyzed. Finally, the group of cycles modulo the closure of the group of boundaries gives (following Lefschetz) a homology group which is always topological; for this we derive a corresponding form of the fundamental theorem. Furthermore, the standard duality between homology and cohomology groups enables us to deduce a corresponding theorem (Theorem 33.1) for the cohomology groups with coefficients in an arbitrary discrete group G .

The fundamental theorem expresses a homology group by means of a group of homomorphisms and a group of group extensions; the latter can also be represented by groups of homomorphisms, as in the basic theorem of Chapter II. The requisite connection between cycles of the homology group and homomorphisms is provided by the Kronecker index (§29).

26. Complexes

The complexes considered here will be abstract cell complexes²⁶ satisfying a star finiteness condition. More precisely, we consider a collection K of abstract elements σ^q called *cells*. With each cell there is associated an integer q called the dimension of σ^q . (There is no restriction requiring the dimension to be non-negative.) To any two cells $\sigma_i^{q+1}, \sigma_j^q$ there corresponds an integer $[\sigma_i^{q+1}:\sigma_j^q]$, called the *incidence number*. K will be called a *star finite complex* provided the incidence numbers satisfy the following two conditions:

(26.1) Given $\sigma_j^q, [\sigma_i^{q+1}:\sigma_j^q] \neq 0$ only for a finite number of indices i ;

(26.2) Given σ_j^{q+1} and σ_k^{q-1} , $\sum_i [\sigma_j^{q+1}:\sigma_i^q][\sigma_i^q:\sigma_k^{q-1}] = 0$.

Condition (26.1) is the star finiteness condition. It insures that the summation in (26.2) is finite.

If we consider the "incidence" matrices of integers

$$A^q = || [\sigma_j^{q+1}:\sigma_i^q] ||$$

we can rewrite the two conditions as follows:

(26.1') A^q is column finite;

(26.2') $A^q A^{q-1} = 0$.

Actually we could have defined a complex as a collection of matrices $\{A^q\}$, $q = 0, \pm 1, \pm 2, \dots$, such that (26.1') and (26.2') hold; we must assume then that the columns of A^q have the same set of labels as do the rows of A^{q-1} , in

²⁶ Essentially like those introduced by A. W. Tucker, for the case of finite complexes. Homology and cohomology are treated as in Whitney [14].

order to form the product $A^q A^{q-1}$. A q -cell will be then either a column of A^q or the corresponding row of A^{q-1} .

A subset L of the cells of K is called an *open subcomplex* if L contains with each q -cell all incident $(q+1)$ -cells; that is, if σ_i^q in L and $[\sigma_i^{q+1}; \sigma_j^q] \neq 0$ imply $\sigma_j^{q+1} \in L$. The incidence matrix A_L^q of L is then the submatrix obtained from A^q by deleting all rows and all columns belonging to cells not in L . Conditions (26.1) and (26.2) automatically hold in L , the latter because of the requirement that L be "open."

A subset $L \subset K$ is a *closed subcomplex* if L contains with each q -cell all incident $(q-1)$ -cells; that is, if $\sigma_i^q \in L$ and $[\sigma_i^q; \sigma_k^{q-1}] \neq 0$ imply $\sigma_k^{q-1} \in L$. The incidence matrix of L is obtained as before, and the conditions (26.1) and (26.2) again hold in L . Whenever L is a closed subcomplex, its complement $K - L$ is an open one, and vice versa.

A subset L of K will be called *q -finite* if L contains only a finite number of q -cells. Because K is star-finite, every $(q-1)$ -cell is contained in a q -finite open subcomplex of K .

27. Homology and cohomology groups

Let G be an abelian group. A q -dimensional *chain* c^q in K with coefficients in G is a function which associates to every q -cell σ_i^q in K an element g_i of G . We write c^q as a formal infinite sum

$$c^q = \sum_i g_i \sigma_i^q.$$

The sum of two chains $\sum g_i \sigma_i^q$ and $\sum h_i \sigma_i^q$ is the chain $\sum (g_i + h_i) \sigma_i^q$, and the chains form a group denoted by $C^q(K, G)$. If $g_i \neq 0$ for only a finite number of indices i then the chain c^q is *finite*. The finite chains form a subgroup $\mathcal{C}_q(K, G)$ of C^q .

The *coboundary* δc^q of a finite chain $c^q = \sum g_i \sigma_i^q$ is defined as

$$\delta c^q = \sum_i \left(\sum_j [\sigma_i^{q+1}; \sigma_j^q] g_j \right) \sigma_i^{q+1}.$$

Because of (26.1) δc^q is a finite $(q+1)$ -chain, while, because of (26.2), $\delta \delta c^q = 0$. The operation δ is a homomorphic mapping of \mathcal{C}_q into \mathcal{C}_{q+1} . The kernel of this homomorphism is a subgroup $\mathcal{Z}_q(K, G)$ of \mathcal{C}_q . The chains of \mathcal{Z}_q are called (finite) *cocycles*:

$$\mathcal{Z}_q(K, G) = [\text{all finite } q\text{-chains } c^q \text{ with } \delta c^q = 0].$$

A *coboundary* is a q -chain of the form δd^{q-1} for some $d^{q-1} \in \mathcal{C}_{q-1}$; these coboundaries form a subgroup

$$\mathcal{B}_q(K, G) = [\text{all finite chains } \delta d^{q-1}].$$

From the relation $\delta \delta = 0$ it follows that $\mathcal{B}_q \subset \mathcal{Z}_q$. The corresponding factor group

$$\mathcal{H}_q(K, G) = \mathcal{Z}_q(K, G) / \mathcal{B}_q(K, G)$$

is called the q^{th} cohomology group of finite cocycles of K with coefficients in G . We also define the co-torsion group $\mathcal{T}_q(K, G)$ as the subgroup of all elements of finite order in $\mathcal{H}_q(K, G)$.

For a chain $c^q = \sum g_i \sigma_i^q$ of $C^q(K, G)$ we also define the boundary

$$\partial c^q = \sum_j \left(\sum_i [\sigma_i^q : \sigma_j^{q-1}] g_i \right) \sigma_j^{q-1}.$$

It again follows from (26.1) that ∂c^q is a well defined chain of $C^{q-1}(K, G)$ and from (26.2) that $\partial \partial c^q = 0$. The operation ∂ is a homomorphic mapping of C^q into C^{q-1} . The kernel of this homomorphism is a subgroup $Z^q(K, G)$ of C^q . The chains of Z^q are called *cycles*:

$$Z^q(K, G) = [\text{all chains } c^q \text{ with } \partial c^q = 0].$$

The chains of the form ∂d^{q+1} where $d^{q+1} \in C^{q+1}$ are the *boundaries*. They form a subgroup

$$B^q(K, G) = [\text{all chains } c^q = \partial d^{q+1}].$$

Because $\partial \partial = 0$ it follows that $B^q \subset Z^q$. The group

$$H^q(K, G) = Z^q(K, G) / B^q(K, G)$$

is called the q^{th} homology group of K with coefficients in G .

Let L be a (closed or open) subcomplex of K . Each chain c^q in K , considered as a function on the q -cells, defines a corresponding chain c_L^q in L . If $c^q = \sum g_i \sigma_i^q$, $c_L^q = \sum' g_i \sigma_i^q$ is the sum found by deleting all terms $g_i \sigma_i^q$ for which σ_i^q is not in L . If L is open, then $\partial_L(c_L^q) = (\partial c^q)_L$, so that one can establish the following facts.

LEMMA 27.1. $c^q \in Z^q(K, G)$ if and only if $c_L^q \in Z^q(L, G)$ for every q -finite open subcomplex L of K .

LEMMA 27.2. If $c^q \in B^q(K, G)$ then $c_L^q \in B^q(L, G)$, provided L is an open subcomplex of K .

A statement analogous to Lemma 27.1 concerning B^q is not generally true. In this connection we define the group $B_w^q(K, G)$ of the *weak boundaries* as follows: $c^q \in B_w^q(K, G)$ provided $c_L^q \in B^q(L, G)$ for every q -finite open subcomplex L of K . For each such open subcomplex L we can construct a subcomplex L' consisting of all q -cells of L , all those $(q+1)$ -cells of L which lie on coboundaries of q -cells of L , and all $(q+i)$ -cells of K , for $i > 1$. This subcomplex L' is open, is both q and $(q+1)$ -finite, and has $B^q(L, G) = B^q(L', G)$. Hence we conclude that $c^q \in B_w^q(K, G)$ if and only if $c_L^q \in B^q(L, G)$ for every open subcomplex L of K which is both q - and $(q+1)$ -finite. Clearly $B^q = B_w^q$ when K itself is q -finite.

It follows from Lemmas 27.1 and 27.2 that

$$B^q(K, G) \subset B_w^q(K, G) \subset Z^q(K, G).$$

The factor group

$$H_w^q(K, G) = Z^q(K, G)/B_w^q(K, G)$$

will be called the *weak q^{th} homology group* of K with coefficients in G . Clearly $H^q = H_w^q$ when K is q -finite.

LEMMA 27.3. $c^q \in B_w^q(K, G)$ if and only if for each finite subset M of K there is a chain c_1^q in $K - M$ such that $c^q - c_1^q \in B^q(K, G)$.

PROOF. Suppose that $c^q \in B_w^q$. Given the finite set M there is a q -finite open subcomplex L containing M . Since $c_L^q \in B^q(L, G)$ there is a d^{q+1} in L such that $(\partial d^{q+1})_L = c_L^q$. Set $c_1^q = c^q - \partial d^{q+1}$. Clearly $c^q - c_1^q \in B^q$ and $(c_1^q)_L = c_L^q - (\partial d^{q+1})_L = 0$, hence $c_1^q \subset K - L \subset K - M$.

Suppose now that c^q satisfies the condition of Lemma 27.3. Given a q -finite open subcomplex L of K there is a c_1^q in $K - L$ such that $c^q - c_1^q \in B^q(K, G)$. There is then a d^{q+1} such that $\partial d^{q+1} = c^q - c_1^q$. Since L is open we have

$$\partial_L(d_L^{q+1}) = (\partial d^{q+1})_L = c_L^q - (c_1^q)_L = c_L^q;$$

therefore $c_L^q \in B^q(L, G)$.

28. Topology in the homology groups

The group of q -chains $C^q(K, G)$ is isomorphic with $\prod_i G_i$, where $G = G_i$ and the set of indices i is in a 1-1 correspondence with the set of q -cells σ_i^q . Hence, if G is a generalized topological group, we can consider $C^q(K, G)$ as a generalized topological group, under the direct product topology, as defined in §1. If G is topological or compact, then $C^q(K, G)$ is also topological or compact, as the case may be.

The boundary operator ∂ , regarded as a homomorphism of C^q into C^{q-1} , is continuous. Since Z^q is the group mapped into 0 by ∂ , we obtain

LEMMA 28.1. If G is topological then $Z^q(K, G)$ is a closed subgroup of $C^q(K, G)$.

From Lemma 27.3 we deduce

LEMMA 28.2. $B^q(K, G) \subset B_w^q(K, G) \subset \bar{B}^q(K, G)$.

The homology groups $H^q = Z^q/B^q$ and $H_w^q = Z^q/B_w^q$ as factor groups of generalized topological groups are generalized topological groups; this is the way they will be considered in the rest of this paper. Even in the case when G and consequently Z^q is topological the groups H^q and H_w^q may be only generalized topological groups, for B^q and B_w^q need not be closed subgroups of Z^q .

If G is compact and topological, then $Z^q(K, G)$ and $C^{q+1}(K, G)$ are compact; since $B^q(K, G)$ is a continuous image of C^{q+1} (under the operation ∂), $B^q(K, G)$ is compact and therefore closed (see Lemma 1.1). Consequently we obtain

LEMMA 28.3. If G is compact and topological, then $B^q(K, G) = B_w^q(K, G) = \bar{B}^q(K, G)$, $H^q(K, G) = H_w^q(K, G)$ and the groups are all compact and topological.

Despite the fact that C_q is a subgroup of the generalized topological group C^q we consider C_q discrete and consequently the cohomology groups $H_q(K, G)$ are taken discrete.

29. The Kronecker index

Let G be a generalized topological group, H a discrete group and assume that a product $\phi(g, h) \in J$ is given pairing G and H to a group J (see §13).

Given two chains

$$c^q \in C^q(K, G), \quad d^q \in \mathcal{C}_q(K, H),$$

we define the Kronecker index as

$$c^q \cdot d^q = \sum_i \phi(g_i, h_i) \in J;$$

the summation is finite since d^q is a finite chain. We verify at once that in this way the groups $C^q(K, G)$ and $\mathcal{C}_q(K, H)$ are paired to J .

Given $c^{q+1} \in C^{q+1}(K, G)$ and $d^q \in \mathcal{C}_q(K, H)$ we have

$$(29.1) \quad (\partial c^{q+1}) \cdot d^q = c^{q+1} \cdot (\delta d^q).$$

This is a restatement of the associative law for matrix multiplication, since the operator ∂ is essentially a postmultiplication by the incidence matrix, while the coboundary operator δ is a premultiplication by the same matrix.

We now examine the annihilators relative to the Kronecker index.

$$(29.2) \quad \mathcal{Z}_q(K, H) \subset \text{Annih } B_w^q(K, G) \subset \text{Annih } B^q(K, G).$$

$$(29.3) \quad Z^q(K, G) \subset \text{Annih } \mathcal{B}_q(K, H).$$

PROOF. Let $z^q \in B_w^q$ and $w^q \in \mathcal{Z}_q$. Since w^q is finite there is a finite subset M of K such that $w^q \subset M$. In view of Lemma 27.3 there is a cycle $z_1^q \subset K - M$ and a chain $c^{q+1} \in C^{q+1}(K, G)$ such that $\partial c^{q+1} = z^q - z_1^q$. Consequently

$$z^q \cdot w^q = (z^q - z_1^q) \cdot w^q = (\partial c^{q+1}) \cdot w^q = c^{q+1} \cdot \delta w^q = c^{q+1} \cdot 0 = 0.$$

Therefore $\mathcal{Z}_q \subset \text{Annih } B_w^q$. The proof of (29.3) is analogous.

It follows from (29.2) and (29.3) that

$$(29.4) \quad H^q(K, G) \text{ and } \mathcal{K}_q(K, H) \text{ are paired to } J,$$

$$(29.5) \quad H_w^q(K, G) \text{ and } \mathcal{K}_q(K, H) \text{ are paired to } J.$$

LEMMA 29.1. *If G and H are dually paired to J then, relative to the Kronecker index,*

$$(29.6) \quad C^q(K, G) \text{ and } \mathcal{C}_q(K, H) \text{ are dually paired to } J,$$

$$(29.7) \quad \mathcal{Z}_q(K, H) = \text{Annih } B_w^q(K, G) = \text{Annih } B^q(K, G),$$

$$(29.8) \quad Z^q(K, G) = \text{Annih } \mathcal{B}_q(K, H).$$

PROOF. Given $c^q = \sum g_i \sigma_i^q \neq 0$ in C^q , we have $g_{i_0} \neq 0$ for some i_0 . Select $h \in H$ so that $\phi(g_{i_0}, h) \neq 0$. Consider the chain $d^q = h \sigma_{i_0}^q$. Then $c^q \cdot d^q = \phi(g_{i_0}, h) \neq 0$. This proves that $\text{Annih } \mathcal{C}_q(K, H) = 0$. Similarly we prove that $\text{Annih } C^q(K, G) = 0$. This establishes (29.6).

Let $d^q \in \text{Annih } B^q(K, G)$. Hence $c^{q+1} \cdot (\delta d^q) = (\partial c^{q+1}) \cdot d^q = 0$ for every c^{q+1} , and therefore $\delta d^q = 0$, in view of (29.6). This shows that $\text{Annih } B^q \subset \mathcal{Z}_q$, which, together with (29.2), gives (29.7).

The proof of (29.8) is analogous to the previous one.

We remark that even when the pairing of the coefficient groups G and H is dual, the pairing (29.4) or (29.5) of the homology and cohomology groups need not be dual, as observed by Whitney ([14], p. 42).

We shall be especially interested in the pairing of G with the group I of integers to G by means of the product $\phi(g, m) = mg$. This pairing has the property that $\text{Annih } I = 0$. This is half of the definition of a dual pairing; the other half ($\text{Annih } G = 0$) may fail in case the order of every element in G divides a fixed integer m . Nevertheless the argument for Lemma 29.1 shows in this case that

$$(29.6') \quad \text{Annih } \mathcal{C}_q(K, I) = 0,$$

$$(29.8') \quad Z^q(K, G) = \text{Annih } \mathcal{B}_q(K, I).$$

We now introduce a subgroup of the group of cycles by the following definition:

$$(29.9) \quad A^q(K, G) = \text{Annih } \mathcal{Z}_q(K, I);$$

in other words, $c^q \in A^q(K, G)$ if and only if $c^q \cdot w^q = 0$ for every finite integral cocycle w^q . The position of this group A^q may be described as follows:

$$(29.10) \quad B_w^q(K, G) \subset A^q(K, G) \subset Z^q(K, G).$$

By (29.2) we have $\mathcal{Z}_q \subset \text{Annih } B_w^q$; consequently $B_w^q \subset \text{Annih } \mathcal{Z}_q = A^q$. Since $\mathcal{B}_q \subset \mathcal{Z}_q$, we have $A^q = \text{Annih } \mathcal{Z}_q \subset \text{Annih } \mathcal{B}_q = Z^q$ by (29.8').

LEMMA 29.2. *If G is a topological group, $A^q(K, G)$ is closed.*

This follows immediately from the continuity of the Kronecker index.

In case G is topological, the various subgroups of cycles of $C^q(K, G)$ are therefore related as follows:

$$B^q \subset B_w^q \subset \bar{B}^q \subset A^q = \bar{A}^q \subset Z^q = \bar{Z}^q \subset C^q.$$

30. Construction of homomorphisms

The essential device of this chapter is that of using the Kronecker index to generate homomorphisms. For a given chain $c^q \in C^q(K, G)$ define θ_{c^q} by

$$(30.1) \quad \theta_{c^q}(d^q) = c^q \cdot d^q, \quad d^q \in \mathcal{C}_q(K, I).$$

LEMMA 30.1. *The correspondence $c^q \rightarrow \theta_{c^q}$ establishes an isomorphism*

$$C^q(K, G) \cong \text{Hom } \{\mathcal{C}_q(K, I), G\}.$$

PROOF. It is clear that $\theta_{c^q} \in \text{Hom } \{\mathcal{C}_q, G\}$, and that the correspondence $c^q \rightarrow \theta_{c^q}$ preserves sums. Also, if $\theta_{c^q} = 0$ then $c^q \cdot d^q = 0$ for all $d^q \in \mathcal{C}_q$ and consequently $c^q = 0$. Conversely, given $\theta \in \text{Hom } \{\mathcal{C}_q, G\}$, define

$$(30.2) \quad c^q = \sum \theta(\sigma_i^q) \sigma_i^q.$$

Clearly $c^q \in C^q(K, G)$, while, for any given $d^q = \sum h_i \sigma_i^q \in \mathcal{C}_q$, we have

$$\theta_{c^q}(d^q) = c^q \cdot d^q = \sum h_i \theta(\sigma_i^q) = \theta(\sum h_i \sigma_i^q) = \theta(d^q).$$

This establishes the algebraic part of the Lemma.

We now recall that

$$C^q(K, G) \cong \prod_i G_i$$

where $G_i = G$ and the subscripts i are in a 1-1 correspondence with the q -cells σ_i^q . On the other hand, since the $\{\sigma_i^q\}$ constitute a set of generators for $\mathcal{C}_q(K, I)$, we have

$$\text{Hom } \{\mathcal{C}_q(K, I), G\} \cong \prod_i G_i.$$

Both these isomorphisms are bicontinuous, hence the combined isomorphism, which is precisely the isomorphism $c^q \leftrightarrow \theta_{c^q}$, is also bicontinuous.

LEMMA 30.2. $\mathcal{Z}_q(K, I)$ is a direct factor of $\mathcal{C}_q(K, I)$.

PROOF. The coboundary operator δ maps \mathcal{C}_q onto \mathcal{B}_{q+1} and the kernel is \mathcal{Z}_q . Hence \mathcal{C}_q is a group extension of \mathcal{Z}_q by \mathcal{B}_{q+1} . As a subgroup of the free group \mathcal{C}_{q+1} the group \mathcal{B}_{q+1} is free (Lemma 4.1) and therefore the group extension is trivial (Theorem 7.2). Hence \mathcal{C}_q is the direct product of \mathcal{Z}_q and a subgroup isomorphic with \mathcal{B}_{q+1} .

THEOREM 30.3. $A^q(K, G)$ is a direct factor of $Z^q(K, G)$ and of $C^q(K, G)$.

PROOF. Since $A^q \subset Z^q$ it will be sufficient to show that A^q is a direct factor of C^q . In the group $\text{Hom } \{\mathcal{C}_q(K, I), G\}$ consider the subgroup A of those homomorphisms that annihilate Z_q . Since \mathcal{Z}_q is a direct factor of \mathcal{C}_q , A is a direct factor of $\text{Hom } \{\mathcal{C}_q, G\}$. However, under the isomorphism $\theta_{c^q} \rightarrow c^q$ of Lemma 30.1 the group A is mapped onto $A^q(K, G) = \text{Annih } \mathcal{Z}_q$, hence the conclusion. This proof also shows (Lemma 3.3) that

$$(30.3) \quad A^q(K, G) \cong \text{Hom } \{\mathcal{C}_q(K, I)/\mathcal{Z}_q(K, I), G\}.$$

Theorem 30.3 leads to the following direct product decompositions of the homology groups:

$$(30.4) \quad H^q(K, G) \cong (Z^q/A^q) \times (A^q/B^q),$$

$$(30.5) \quad H_w^q(K, G) \cong (Z^q/A^q) \times (A^q/B_w^q).$$

We proceed with the study of the first factor, Z^q/A^q .

THEOREM 30.4. The correspondence $c^q \rightarrow \theta_{c^q}$ establishes an isomorphism

$$Z^q(K, G)/A^q(K, G) \cong \text{Hom } \{\mathcal{K}_q(K, I), G\}.$$

PROOF. Since $Z^q = \text{Annih } \mathcal{B}_q$, by (29.8'), it follows that under the isomorphism $c^q \rightarrow \theta_{c^q}$ the group Z^q is mapped onto the subgroup of $\text{Hom } \{\mathcal{C}_q, G\}$ consisting of those homomorphisms annihilating \mathcal{B}_q . By Lemma 3.3 the latter subgroup can be identified with $\text{Hom } \{\mathcal{C}_q/\mathcal{B}_q, G\}$, so $Z^q \cong \text{Hom } \{\mathcal{C}_q/\mathcal{B}_q, G\}$. On the other hand, $\mathcal{Z}_q/\mathcal{B}_q$ is a direct factor of $\mathcal{C}_q/\mathcal{B}_q$, so that Lemma 3.4 shows

that $\text{Hom } \{\mathcal{Z}_q/\mathcal{B}_q, G\}$ is a factor group of $\text{Hom } \{\mathcal{C}_q/\mathcal{B}_q, G\}$, corresponding to the subgroup consisting of homomorphisms annihilating $\mathcal{Z}_q/\mathcal{B}_q$. This subgroup in turn corresponds to the subgroup A^q of Z^q , hence

$$Z^q/A^q \cong \text{Hom } \{\mathcal{Z}_q/\mathcal{B}_q, G\}.$$

This is the desired conclusion.

31. Study of A^q

The correspondence $c^q \rightarrow \theta_{c^q}$ of Lemma 30.1 maps the group A^q of annihilators of cocycles onto the group of those homomorphisms of \mathcal{C}_q into G which carry \mathcal{Z}_q into zero. As observed in Lemma 3.3, the latter group is isomorphic to $\text{Hom } \{\mathcal{C}_q/\mathcal{Z}_q, G\}$. Since $\mathcal{C}_q/\mathcal{Z}_q \cong \mathcal{B}_{q+1}$, this gives the isomorphism

$$(31.1) \quad A^q(K, G) \cong \text{Hom } \{\mathcal{B}_{q+1}(K, I), G\}.$$

An examination of this construction shows that the homomorphism corresponding to a given $z^q \in A^q$ is determined as follows. For each $d^{q+1} \in \mathcal{B}_{q+1}$ choose a $d^q \in \mathcal{C}_q(K, I)$ for which $\delta d^q = d^{q+1}$, and define²⁷

$$\phi_{z^q}(d^{q+1}) = z^q \cdot d^q.$$

Because z^q is in A^q , this result is independent of the choice of d^q for given d^{q+1} . Furthermore ϕ_{z^q} is a homomorphism of \mathcal{B}_{q+1} into G , and it is obtained from θ_{z^q} by the process indicated above, for one has

$$\phi_{z^q}(\delta d^q) = \theta_{z^q}(d^q).$$

We therefore have the following result.

LEMMA 31.1. *The correspondence $z^q \rightarrow \phi_{z^q}$ establishes the (bicontinuous) isomorphism (31.1).*

The properties of this isomorphism can be collected in the following

THEOREM 31.2. *The isomorphism $z^q \rightarrow \phi_{z^q}$ induces the isomorphisms*

$$A^q(K, G)/B^q(K, G) \cong \text{Hom } \{\mathcal{B}_{q+1}, G\}/\text{Hom } \{\mathcal{Z}_{q+1} \mid \mathcal{B}_{q+1}, G\},$$

$$B_w^q(K, G)/B^q(K, G) \cong \text{Hom}_f \{\mathcal{B}_{q+1}, G; \mathcal{Z}_{q+1}\}/\text{Hom } \{\mathcal{Z}_{q+1} \mid \mathcal{B}_{q+1}, G\},$$

$$A^q(K, G)/B_w^q(K, G) \cong \text{Hom } \{\mathcal{B}_{q+1}, G\}/\text{Hom}_f \{\mathcal{B}_{q+1}, G; \mathcal{Z}_{q+1}\},$$

where $\mathcal{B}_{q+1} = \mathcal{B}_{q+1}(K, I)$ and $\mathcal{Z}_{q+1} = \mathcal{Z}_{q+1}(K, I)$.

PROOF. We shall show that the groups $B^q(K, G)$ and $B_w^q(K, G)$ are mapped onto $\text{Hom } \{\mathcal{Z}_{q+1} \mid \mathcal{B}_{q+1}, G\}$ and $\text{Hom}_f \{\mathcal{B}_{q+1}, G; \mathcal{Z}_{q+1}\}$, respectively.

Assume that $z^q \in B^q(K, G)$; then $\partial z^{q+1} = z^q$ for some $z^{q+1} \in C^{q+1}(K, G)$. Define

$$\phi^*(d^{q+1}) = z^{q+1} \cdot d^{q+1}; \quad d^{q+1} \in C_{q+1}.$$

²⁷ Notice the analogy with the definition of the so-called "linking coefficient" (cf. Lefschetz [7], Ch. III).

Clearly $\phi^* \in \text{Hom} \{C_{q+1}, G\}$. If $d^{q+1} = \delta d^q$ then

$$\phi^*(d^{q+1}) = z^{q+1} \cdot d^{q+1} = z^{q+1} \cdot \delta d^q = \partial z^{q+1} \cdot d^q = z^q \cdot d^q = \phi_{x^q}(d^{q+1}).$$

Hence ϕ^* is an extension of ϕ_{x^q} to C_{q+1} and in particular also to Z_{q+1} .

Suppose conversely that ϕ_{x^q} can be extended to Z_{q+1} . Since Z_{q+1} is a direct factor of C_{q+1} (Lemma 30.2) we may then find an extension ϕ^* of ϕ_{x^q} to C_{q+1} . Define

$$z^{q+1} = \sum_i \phi^*(\sigma_i^{q+1}) \sigma_i^{q+1}.$$

Clearly $z^{q+1} \in C^{q+1}(K, G)$ and $z^{q+1} \cdot \sigma_j^{q+1} = \phi^*(\sigma_j^{q+1})$ and hence $z^{q+1} \cdot d^{q+1} = \phi^*(d^{q+1})$ for all $d^{q+1} \in C_{q+1}$. Consequently

$$\partial z^{q+1} \cdot \sigma_j^q = z^{q+1} \cdot \delta \sigma_j^q = \phi^*(\delta \sigma_j^q) = \phi_{x^q}(\delta \sigma_j^q) = z^q \cdot \sigma_j^q.$$

Since this holds for every σ_j^q we have $\partial z^{q+1} = z^q \in B^q(K, G)$.

Suppose $z^q \in B_w^q(K, G)$. In view of Lemma 5.1 it is sufficient to prove that if the cocycle $md^{q+1} \in \mathcal{B}_{q+1}(K)$ then $\phi_{x^q}(md^{q+1})$ is divisible by m . Let $\delta d^q = md^{q+1}$ and let M be a finite subset of K such that $d^q \subset M$. In view of Lemma 27.3 there is a chain $z_1^q \subset K - M$ such that $z^q - z_1^q = z_2^q \in B^q(K, G)$. It follows that $z_2^q \in A^q$ and so that $z_1^q \in A^q$, hence $\phi_{x_1^q}$ and $\phi_{x_2^q}$ are defined and $\phi_{x^q} = \phi_{x_1^q} + \phi_{x_2^q}$. Since $z_2^q \in B^q(K, G)$, then, as we just proved, $\phi_{x_2^q}$ can be extended to Z_{q+1} and therefore $\phi_{x_2^q}(md^{q+1})$ must be divisible by m . Since $d^q \subset M$ and $z_1^q \subset K - M$ we have $\phi_{x_1^q}(md^{q+1}) = z_1^q \cdot d^q = 0$. Hence $\phi_{x^q}(md^{q+1})$ is divisible by m .

Suppose conversely that ϕ_{x^q} can be extended to every subgroup of $Z_{q+1}(K, I)$ of finite order over $\mathcal{B}_{q+1}(K, I)$. Then, as in Lemma 5.2, ϕ_{x^q} can also be extended to every subgroup \mathcal{D} of $Z_{q+1}(K, I)$ such that $\mathcal{D}/\mathcal{B}_{q+1}$ has a finite number of generators. Now let L be any open subcomplex of K which is both q and $(q+1)$ -finite; there is then an extension of ϕ_{x^q} to the group \mathcal{D}_L generated by $\mathcal{B}_{q+1}(K, I)$ and $Z_{q+1}(L, I)$. But in the complex L the homomorphism ϕ_{y^q} induced by $y^q = z_L^q$ agrees on $\mathcal{B}_{q+1}(L, I)$ with the homomorphism ϕ_{x^q} . Therefore $\phi_{y^q} \in \text{Hom} \{\mathcal{B}_{q+1}(L, I), G\}$ has an extension to $Z_{q+1}(L, I)$. In view of what we proved before, we therefore have $y^q = z_L^q \in B^q(L, G)$. Since this holds for each L considered, $z^q \in B_w^q(K, G)$. This concludes the proof of Theorem 31.2.

In this theorem the factor homomorphism groups on the right can be reinterpreted as groups of group extensions, in accord with the results of Chapter II.

THEOREM 31.3. *The isomorphism $z^q \leftrightarrow \phi_{x^q}$ combined with the isomorphisms establishing relations between group extensions and homomorphisms lead to the following isomorphisms:*

$$(31.2) \quad A^q(K, G)/B^q(K, G) \cong \text{Ext} \{G, \mathcal{K}_{q+1}\},$$

$$(31.3) \quad B_w^q(K, G)/B^q(K, G) \cong \text{Ext}_f \{G, \mathcal{K}_{q+1}\},$$

$$(31.4) \quad A^q(K, G)/B_w^q(K, G) \cong \text{Ext} \{G, \mathcal{T}_{q+1}\}/\text{Ext}_f \{G, \mathcal{T}_{q+1}\}$$

where $\mathcal{K}_{q+1} = \mathcal{K}_{q+1}(K, I)$ and $\mathcal{T}_{q+1} = \mathcal{T}_{q+1}(K, I)$ is the corresponding co-torsion group.

The isomorphisms established so far have all been bicontinuous.

32. Computation of the homology groups

As we have shown in §29, the Kronecker index establishes a pairing of the group $H^q(K, G)$ or $H_w^q(K, G)$ with the group $\mathcal{K}_q(K, I)$, the values of the products being in the group G . Accordingly we define the following subhomology groups:

$$(32.1) \quad Q^q(K, G) = \text{Annih } \mathcal{K}^q(K, I) \text{ in } H^q(K, G),$$

$$(32.2) \quad Q_w^q(K, G) = \text{Annih } \mathcal{K}_q(K, I) \text{ in } H_w^q(K, G).$$

We verify at once that $Q^q = A^q/B^q$ and $Q_w^q = A^q/B_w^q$. Consequently the results of the last two sections furnish the following two basic theorems:

THEOREM 32.1. *For a star finite complex K the homology group $H^q(K, G)$ of infinite cycles with coefficients in a generalized topological group G can be expressed in terms of the integral cohomology groups $\mathcal{K}_q = \mathcal{K}_q(K, I)$ and $\mathcal{K}_{q+1} = \mathcal{K}_{q+1}(K, I)$ of finite cocycles. The explicit relation is*

$$(32.3) \quad H^q(K, G) \cong \text{Hom } \{\mathcal{K}_q, G\} \times \text{Ext } \{G, \mathcal{K}_{q+1}\}.$$

More explicitly, H^q has a subgroup Q^q , defined by (32.1), where

$$(32.4) \quad Q^q(K, G) \text{ is a direct factor of } H^q(K, G),$$

$$(32.5) \quad Q^q(K, G) \cong \text{Ext } \{G, \mathcal{K}_{q+1}\},$$

$$(32.6) \quad H^q(K, G)/Q^q(K, G) \cong \text{Hom } \{\mathcal{K}_q, G\}.$$

THEOREM 32.2. *For a star finite complex K the weak homology group $H_w^q(K, G)$ of infinite cycles with coefficients in a generalized topological group G can be expressed in terms of the integral cohomology group $\mathcal{K}_q = \mathcal{K}_q(K, I)$ and the integral co-torsion group $\mathcal{I}_{q+1} = \mathcal{I}_{q+1}(K, I)$ of finite cocycles. The explicit relation is*

$$(32.7) \quad H_w^q(K, G) \cong \text{Hom } \{\mathcal{K}_q, G\} \times (\text{Ext } \{G, \mathcal{I}_{q+1}\} / \text{Ext}_f \{G, \mathcal{I}_{q+1}\}).$$

More explicitly, H_w^q has a subgroup Q_w^q , defined by (32.2), where

$$(32.8) \quad Q_w^q(K, G) \text{ is a direct factor of } H_w^q(K, G),$$

$$(32.9) \quad Q_w^q(K, G) \cong \text{Ext } \{G, \mathcal{I}_{q+1}\} / \text{Ext}_f \{G, \mathcal{I}_{q+1}\},$$

$$(32.10) \quad H_w^q(K, G)/Q_w^q(K, G) \cong \text{Hom } \{\mathcal{K}_q, G\}.$$

Both factors in (32.3) and (32.7) are generalized topological groups and the isomorphisms are bicontinuous.

If G is topological then by Corollary 3.2 the group $\text{Hom } \{\mathcal{K}_q, G\}$ is topological. If we also assume that mG is a closed subgroup of G for $m = 2, 3, \dots$ then Corollary 11.6 shows that $\text{Ext}_f \{G, \mathcal{I}_{q+1}\}$ is a closed subgroup of $\text{Ext } \{G, \mathcal{I}_{q+1}\}$. Consequently we obtain

THEOREM 32.3. (Steenrod [9]). *If G is a topological group and mG is a closed subgroup of G for $m = 2, 3, \dots$ then $H_w^q(K, G)$ is topological.*

The expressions for Q^q and Q_w^q can be simplified if additional information concerning the group G is available. If G is infinitely divisible then, by Corollary 11.4, $\text{Ext } \{G, H\} = 0$ for all H and therefore

COROLLARY 32.4. *If G is infinitely divisible then $Q^q(K, G) = Q_w^q(K, G) = 0$ and*

$$H^q(K, G) = H_w^q(K, G) \cong \text{Hom } \{\mathcal{K}_q, G\}.$$

From Theorem 17.2 we deduce

COROLLARY 32.5. *If G has no elements of finite order then*

$$Q_w^q(K, G) \cong \text{Ext } \{G, \mathcal{I}_{q+1}\}.$$

If, in addition, G is discrete then

$$Q_w^q(K, G) \cong \text{Hom } \{\mathcal{I}_{q+1}, G_\infty/G\}.$$

In particular, if $G = I$ then, by Theorem 17.1, $Q_w^q(K, I) \cong \text{Char } \mathcal{I}_{q+1}$ and therefore

$$(32.11) \quad H_w^q(K, I) \cong \text{Hom } \{\mathcal{K}_q, I\} \times \text{Char } \mathcal{I}_{q+1}.$$

THEOREM 32.6. *If G is compact and topological then $H^q(K, G) = H_w^q(K, G)$ is compact and topological and*

$$(32.12) \quad Q^q(K, G) = Q_w^q(K, G) \cong \text{Ext } \{G, \mathcal{I}_{q+1}\} \cong \text{Char Hom } \{G, \mathcal{I}_{q+1}\}.$$

This is a consequence of Corollary 11.7 and Theorem 15.1. Since G is compact, \mathcal{I}_{q+1} discrete, and only continuous homomorphisms are taken in $\text{Hom } \{G, \mathcal{I}_{q+1}\}$, it follows that in the formula (32.12) for $Q^q(K, G)$ we may replace G by G/G_0 where G_0 is the component of 0 in G .

COROLLARY 32.7. *If $\mathcal{K}_{q+1}(K, I)$ has a finite number of generators then $B^q(K, G) = B_w^q(K, G)$ and*

$$(32.13) \quad H^q(K, G) = H_w^q(K, G) \cong \text{Hom } \{\mathcal{K}_q, G\} \times \text{Ext } \{G, \mathcal{I}_{q+1}\}.$$

In fact, since $\text{Ext}_f \{G, \mathcal{K}_{q+1}\} = 0$ (Corollary 11.3) it follows from (31.3) that $B^q = B_w^q$. Since also $\text{Ext}_f \{G, \mathcal{I}_{q+1}\} = 0$, formula (32.13) follows from Theorem 32.2.

In particular, Corollary 32.7 applies if K is a finite complex (cf. Alexandroff-Hopf [1], Ch. V and Steenrod [9], p. 675).

33. Computation of the cohomology groups

We start out with a brief review of the duality between homology and cohomology. Let G be a discrete group and $\hat{G} = \text{Char } G$ compact and topological. Since \hat{G} and G are dually paired to the group P of reals mod 1 (see §13) the Kronecker index $c^q \cdot d^q \in P$ is defined as in §29 for $c^q \in C^q(K, \hat{G})$ and $d^q \in \mathcal{C}_q(K, G)$. Since the pairing of \hat{G} and G is dual (Theorem 13.5) we have by Lemma 29.1

$$(33.1) \quad C^q(K, \hat{G}) \text{ and } \mathcal{C}_q(K, G) \text{ are dually paired to } P,$$

$$(33.2) \quad \mathcal{Z}_q(K, G) = \text{Annih } B^q(K, \hat{G}); \quad Z^q(K, \hat{G}) = \text{Annih } \mathcal{B}_q(K, G).$$

These formulas, Theorem 13.7, Lemma 13.2 and Theorem 13.5 imply that the Kronecker index defines a dual pairing of $\mathcal{K}_q(K, G)$ and $H^q(K, \hat{G})$ to P and that

$$(33.3) \quad \mathcal{K}_q(K, G) \cong \text{Char } H^q(K, \text{Char } G).$$

Using this result and the formulas established in the previous section for $H^q(K, \text{Char } G)$ we could write down a formula expressing $\mathcal{K}_q(K, G)$. For convenience we first define a subcohomology group

$$(33.4) \quad \mathcal{P}_q(K, G) = \text{Annih } Q^q(K, \text{Char } G) \text{ in } \mathcal{K}_q(K, G),$$

in order to get a more detailed form for our result.

THEOREM 33.1. *For a star finite complex K the cohomology group $\mathcal{K}_q(K, G)$ of finite cocycles with coefficients in a discrete group G can be expressed in terms of the cohomology group $\mathcal{K}_q = \mathcal{K}_q(K, I)$ and the integral co-torsion group $\mathcal{I}_{q+1} = \mathcal{I}_{q+1}(K, I)$. The explicit relation is*

$$(33.5) \quad \mathcal{K}_q(K, G) \cong (G \circ \mathcal{K}_q) \times \text{Hom } \{\text{Char } G, \mathcal{I}_{q+1}\}.$$

More explicitly, $\mathcal{K}_q(K, G)$ has a subgroup $\mathcal{P}_q(K, G)$, defined by (33.4), where

$$(33.6) \quad \mathcal{P}_q(K, G) \text{ is a direct factor of } \mathcal{K}_q(K, G),$$

$$(33.7) \quad \mathcal{P}_q(K, G) \cong G \circ \mathcal{K}_q$$

$$(33.8) \quad \mathcal{K}_q(K, G) / \mathcal{P}_q(K, G) \cong \text{Hom } \{\text{Char } G, \mathcal{I}_{q+1}\}.$$

PROOF. Since Q^q is a direct factor of H^q it follows from the character theory that $\mathcal{P}_q = \text{Annih } Q^q$ is a direct factor of $\mathcal{K}_q(K, G) = \text{Char } H^q$. It also follows that

$$\mathcal{P}_q \cong \text{Char } (H^q / Q^q), \quad \mathcal{K}_q(K, G) / \mathcal{P}_q(K, G) \cong \text{Char } Q^q.$$

The first formula and (32.6) imply

$$\mathcal{P}_q(K, G) \cong \text{Char Hom } \{\mathcal{K}_q, \text{Char } G\},$$

which in view of Theorem 18.1 gives (33.7). The second formula combined with (32.12) proves (33.8).

If G has no elements of finite order, then $\text{Char } G$ is connected and therefore $\text{Hom } \{\text{Char } G, \mathcal{I}_{q+1}\} = 0$. From (33.7) and (33.8) we therefore obtain

COROLLARY 33.2. *If G has no elements of finite order then*

$$\mathcal{K}_q(K, G) = \mathcal{P}_q(K, G) \cong G \circ \mathcal{K}_q(K, I).$$

We now proceed to give an intrinsic characterization of the subgroup \mathcal{P}_q of $\mathcal{K}_q(K, G)$. A cocycle $w^q \in \mathcal{Z}_q(K, G)$ will be called *pure* if it is a linear combination of integral cocycles, as

$$w_q = \sum_{i=1}^k g_i w_i^q, \quad g_i \in G, \quad w_i^q \in \mathcal{Z}_q(K, I).$$

LEMMA 33.3. *The group $\mathcal{P}_q(K, G)$ is the subgroup of $\mathcal{K}_q(K, G)$ determined by the pure cocycles.*

PROOF. Let \mathcal{S} be the subgroup of $\mathcal{Z}_q(K, G)$ consisting of all the pure cocycles. It may be shown that $\mathcal{B}_q(K, G) \subset \mathcal{S}$. In order to prove that $\mathcal{S} / \mathcal{B}_q(K, G) = \mathcal{P}_q(K, G)$ we must prove that $\mathcal{S} / \mathcal{B}_q(K, G) = \text{Annih } Q^q(K, \hat{G})$ where $\hat{G} = \text{Char } G$.

This is equivalent to proving that $Q^q(K, \hat{G}) = \text{Annih } (\mathfrak{S}/\mathfrak{B}_q(K, G))$, which reduces to the formula

$$A^q(K, \hat{G}) = \text{Annih } \mathfrak{S},$$

that we now propose to establish.

Let $z^q \in A^q(K, \hat{G})$ and let $w^q \in \mathfrak{S}$. Since $w^q = \sum g_i w_i^q$, where $w_i^q \in \mathfrak{Z}_q(K, I)$ and since $z^q \cdot w_i^q = 0$ by the definition of A^q , it follows that $z^q \cdot w^q = 0$.

Suppose now that c^q lies in $C^q(K, \hat{G})$ but not in $A^q(K, \hat{G})$. There is then a $w_i^q \in \mathfrak{Z}_q(K, I)$ such that $c^q \cdot w_i^q = \hat{g} \neq 0$ where $\hat{g} \in \hat{G}$. Pick $g \in G$ so that $\hat{g}(g) \neq 0$ and define $w^q = g w_i^q$. Clearly $w^q \in \mathfrak{S}$ is a pure cocycle and $c^q \cdot w^q = \hat{g}(g) \neq 0$, hence c^q is not in $\text{Annih } \mathfrak{S}$. This concludes the proof of the Lemma.

Using the description of $\mathcal{P}_q(K, G)$ given in the Lemma we could easily establish the isomorphism $\mathcal{P}_q \cong G \circ \mathcal{K}_q(K, I)$ directly, using the definition of the tensor product. This was the procedure adopted by Čech [3] who essentially has proved all the results of this section. Our main improvement is that our isomorphisms are given explicitly and invariantly, while Čech used generators and relations throughout.

34. The groups H_i^q

The fact that the groups H^q and H_w^q may not be topological groups even though the coefficient group G is chosen to be topological induced Lefschetz and others to introduce the following group, for a topological coefficient group G ,

$$H_i^q(K, G) = Z^q(K, G)/B^q(K, G)$$

as a standard homology group for K .

The relation of this group to the groups previously considered is immediate:

$$(34.1) \quad H_i^q \cong H^q/\bar{0} \cong H_w^q/\bar{0}.$$

Theorem 32.3 can now be reformulated as follows.

THEOREM 34.1 (Steenrod [9]) *If G is topological and mG is closed for $m = 2, 3, \dots$ then $H_w^q(K, G) = H_i^q(K, G)$.*

Since G is a topological group, $A^q(K, G)$ is a closed subgroup of $Z^q(K, G)$ (Lemma 29.2) and consequently $B^q \subset A^q$. It follows that the Kronecker index can be defined for elements of $H_i^q(K, G)$ and $\mathcal{K}_q(K, I)$. We define a sub-homology group

$$(34.2) \quad Q_i^q(K, G) = \text{Annih } \mathcal{K}_q(K, I) \text{ in } H_i^q(K, G).$$

THEOREM 34.2. *For a star finite complex K the topological homology group $H_i^q(K, G)$ of infinite cycles with coefficients in a topological group G can be expressed in terms of the integral cohomology group $\mathcal{K}_q = \mathcal{K}_q(K, I)$ and the integral co-torsion group $\mathcal{J}_{q+1} = \mathcal{J}_{q+1}(K, I)$ of finite cocycles. The explicit relation is*

$$(34.3) \quad H_i^q(K, G) \cong \text{Hom } \{\mathcal{K}_q, G\} \times (\text{Ext } \{G, \mathcal{J}_{q+1}\}/\bar{0}).$$

More explicitly, H_i^q has a subgroup Q_i^q , defined by (34.2), where

$$(34.3) \quad Q_i^q(K, G) \text{ is a direct factor of } H_i^q(K, G),$$

$$(34.5) \quad Q_i^q(K, G) \cong \text{Ext} \{G, \mathcal{I}_{q+1}\} / \bar{0},$$

$$(34.6) \quad H_i^q(K, G) / Q_i^q(K, G) \cong \text{Hom} \{\mathcal{K}_q, G\}.$$

PROOF. From the direct product decomposition (30.5) we obtain

$$H_i^q \cong (Z^q / A^q) \times [(A^q / B_w^q) / \bar{0}].$$

Consequently $Q_i^q = Q_w^q / \bar{0}$ is a direct factor. Since $Q_w^q \cong \text{Ext} \{G, \mathcal{I}_{q+1}\} / \text{Ext}_f \{G, \mathcal{I}_{q+1}\}$ and since, by Corollary 11.6, $\overline{\text{Ext}_f \{G, \mathcal{I}_{q+1}\}} = \bar{0}$, we obtain (34.5). Formula (34.6) follows from Theorem 30.4.

It might be interesting to notice that, while the groups $H^q(K, G)$ and $H_w^q(K, G)$ were algebraically independent of the choice of the topology in G , the group $H_i^q(K, G)$ depends both algebraically and topologically upon the topology chosen in G .

35. Universal coefficients

The results of the previous three sections can be summarized in the following fashion.

UNIVERSAL COEFFICIENT THEOREM. *In a star finite complex K the integral cohomology groups of finite cocycles determine all the homology and cohomology groups that were defined for a star finite complex, specifically:*

The groups G , $\mathcal{K}_q(K, I)$ and $\mathcal{K}_{q+1}(K, I)$ determine the generalized topological homology group $H^q(K, G)$ of infinite cycles with coefficients in a generalized topological group G .

The groups G , $\mathcal{K}_q(K, I)$ and $\mathcal{I}_{q+1}(K, I)$ determine:

(a) *the generalized topological weak homology group $H_w^q(K, G)$ of infinite cycles with coefficients in a generalized topological group G ;*

(b) *the topological homology group $H_i^q(K, G)$ of infinite cycles with coefficients in a topological group G ;*

(c) *the discrete cohomology group $\mathcal{K}_q(K, G)$ of finite cocycles with coefficients in a discrete group G .*

This shows that the group I of integers is a universal coefficient group for the homology theory of the complex K . Since the group P of reals mod 1 is the group of characters of I we have in view of (33.3) the fact that $\mathcal{K}_q(K, I) \cong \text{Char } H^q(K, P)$; therefore all the groups can be expressed in terms of $H^q(K, P)$ and $H^{q+1}(K, P)$, so that P is also universal.

Given a closed subcomplex L of K one often has to consider the relative groups of K mod L . However, the complexes used here are so general that $K - L$ is also a complex and the usual groups of K mod L coincide with the groups of $K - L$ as we have defined them. Consequently all our formulas remain valid in the relative theory.

36. Closure finite complexes

Closure finite complexes are obtained by replacing condition (26.1) in the definition of a complex by the following

(36.1) *Given σ_i^q , $[\sigma_i^q: \sigma_k^{q-1}] \neq 0$ for only a finite number of indices k .*

Simplicial complexes are all closure finite.

In a closure finite complex we consider finite cycles and infinite cocycles and obtain the discrete homology groups $\mathcal{H}^q(K, G)$ and the topologized cohomology groups $H_q(K, G)$, $H_q^w(K, G)$ and $H_q^t(K, G)$. All our development can be repeated with the modification of interchanging homology and cohomology groups and replacing $q + 1$ by $q - 1$. For instance formula (32.3) will take the form:

$$H_q(K, G) \cong \text{Hom} \{ \mathcal{H}^q(K, I), G \} \times \text{Ext} \{ G, \mathcal{H}^{q-1}(K, I) \}.$$

Instead of repeating the arguments for closure finite complexes we can use the previous results for star finite complexes and apply them to closure finite complexes by means of the concept of the dual complex. If the complex K is described by the incidence matrices A^q , the dual complex K^* will be defined by the transposed matrices

$$B^q = (A^{-q})'$$

The dual of a star finite complex is closure finite and vice versa. Also $(K^*)^* = K$. Moreover by passing from a complex to its dual, the boundary operation becomes the coboundary, and vice versa. Hence the homology and cohomology group are interchanged, and our formulas apply.

A locally finite (i.e. both closure and star finite) complex carries therefore two homology theories, namely, the theory of a star finite complex and the theory of a closure finite one. In the case of a manifold the Poincaré duality establishes a relation between the two theories. In general the theories are unrelated and in any specific problem we only use one at a time. We will quote two examples to this effect.

A) In the following chapter we define for every compact metric space a complex called the fundamental complex. This complex is locally finite, but its closure finite theory is trivial, while its star finite theory is extremely useful for the study of the underlying space.

B) Let us consider two infinite polyhedra represented as two locally finite complexes K and K' . Given a continuous mapping f of K into K' it is well known that f induces homomorphisms: 1°) of the groups of finite cycles of K into the corresponding groups of K' , 2°) of the groups of infinite cocycles of K' into the corresponding groups of K . This explains why in problems connected with continuous mappings (like Hopf's mapping theorem and its generalizations; see [4]) we use only finite cycles and infinite cocycles, or in other words we use only the closure finite theory of K and K' .

CHAPTER VI. TOPOLOGICAL SPACES

Here we formulate our results for the homology groups of a space. In the case of a compact metric space, Steenrod has shown that the homology groups can all be expressed as corresponding homology groups of the fundamental complex of the space, so that the results of Chapter V apply directly (§44). For a general space, the Čech homology groups are obtained as (direct or inverse) limits, so that the decomposition of the homology group is obtained as a limit of the known decompositions for the homology groups of finite complexes, and here the techniques developed in Chapter IV apply. The results obtained for a general space are not as complete as those for complexes, partly because the limit of a set of direct sums apparently need not be a direct sum, and partly because "Lim" and "Ext" do not permute, so that the group Ext^* discussed in Chap. IV is requisite. We also discuss (§45) Steenrod's homology groups of "regular" cycles.

37. Chain transformations

Let $K = \{\sigma_i^q\}$ and $K' = \{\tau_j^q\}$ be two star finite complexes. Suppose also that for every integer q there is given a matrix of integers,

$$B^q = || b_{ij}^q ||$$

with rows indexed by the q -cells of K , columns by the q -cells of K' , and with only a finite number of non-zero entries in each column.

Given a q -chain $c^q = \sum g_i \sigma_i^q \in C^q(K, G)$ in K , define

$$Tc^q = \sum_j \left(\sum_i g_i b_{ij}^q \right) \tau_j^q.$$

The column finiteness condition implies that the summation $\sum_i g_i b_{ij}^q$ is finite and therefore that Tc^q is a well defined element of $C^q(K', G)$. We thus obtain homomorphisms (one for each q and G)

$$T: C^q(K, G) \rightarrow C^q(K', G).$$

Given a finite q -chain $d^q = \sum g_j \tau_j^q \in C_q(K', G)$ in K' , define

$$T^* d^q = \sum_i \left(\sum_j g_j b_{ij}^q \right) \sigma_i^q$$

This time the column finiteness of B^q implies that $T^* d^q$ is finite; hence we obtain homomorphisms

$$T^*: C_q(K', G) \rightarrow C_q(K, G).$$

T^* is called the *dual* of T .

It can be verified at once that if c^q is a chain in K and d^q is a finite chain in K' then

$$(37.1) \quad (Tc^q) \cdot d^q = c^q \cdot (T^* d^q),$$

whenever the coefficients are such that the Kronecker index has a meaning (§29).

T is called a *chain transformation* of K into K' if $\partial Tc^q = T(\partial c^q)$ for every q chain; that is, if

$$(37.2) \quad \partial T = T\partial.$$

It can be shown that this condition is equivalent to the requirement that

$$(37.3) \quad \delta T^* = T^*\delta.$$

It follows that a chain transformation T maps the groups Z^q , A^q , B_w^q and B^q of K homomorphically into the corresponding groups of K' . Similarly T^* maps the groups of K' into the corresponding groups of K . In particular a chain transformation induces homomorphisms of the homology groups

$$(37.4) \quad T: H^q(K, G) \rightarrow H^q(K', G),$$

$$(37.5) \quad T^*: \mathcal{H}_q(K', G) \rightarrow \mathcal{H}_q(K, G),$$

and of the corresponding subgroups defined by (32.1) and (33.4)

$$(37.6) \quad T: Q^q(K, G) \rightarrow Q^q(K', G),$$

$$(37.7) \quad T^*: \mathcal{P}_q(K', G) \rightarrow \mathcal{P}_q(K, G).$$

38. Naturality

We are now in a position to give a precise meaning to the fact that the isomorphisms established in Chapter V are all "natural."

THEOREM 38.1. *If T is a chain transformation of a complex K into K' , then T permutes with the isomorphisms established in Theorems 30.4 and 31.2, provided the application of T in any group is taken to mean the application of the appropriate transformation induced by T on that group.*

PROOF. If the homomorphism established in Theorem 30.4 be denoted by μ (or by μ' , for K'), then we have the homomorphisms

$$\begin{array}{ccc} Z^q(K) & \xrightarrow{\mu} & \text{Hom } \{\mathcal{H}_q, G\} \\ \downarrow T & & \downarrow T_h^{**} \\ Z^q(K') & \xrightarrow{\mu'} & \text{Hom } \{\mathcal{H}'_q, G\}, \end{array}$$

where T_h^{**} is the homomorphism of $\text{Hom } \{\mathcal{H}_q(K, I), G\}$ into $\text{Hom } \{\mathcal{H}_q(K', I), G\}$, induced as in §12 by the dual chain transformation T^* . The theorem then asserts that

$$\mu'T = T_h^{**}\mu.$$

To show this, take $c^q \in Z^q(K, G)$. The corresponding homomorphism $\theta = \mu c^q$ is then defined, for each cocycle d^q in $\mathcal{Z}_q(K)$, by $\theta(d^q) = c^q \cdot d^q$ (cf. §30). Then $\theta' = T_h^{**}\theta$ is, according to the definition of T_h , simply $\theta'(d'^q) = \theta(T^*d'^q)$. Hence, for any cocycle d'^q ,

$$\theta'(d'^q) = \theta(T^*d'^q) = c^q \cdot (T^*d'^q) = (Tc^q) \cdot d'^q.$$

In the other direction, Tc^q maps under μ' into the homomorphism ϕ' , defined for $d'^q \in Z_q(K')$ by

$$\phi'(d'^q) = (Tc^q) \cdot d'^q.$$

The formulas show that $\phi' = \mu' Tc^q$ and $\theta' = T_h^{**} \mu c^q$ are in fact identical, as required by Theorem 38.1.

To treat Theorem 31.2, let τ (or τ') denote the homomorphism of $A^q(K, G)$ onto $\text{Hom} \{\mathcal{B}_{q+1}(K, I), G\}$ given in that theorem, while η is the map of the latter group onto $\text{Ext} \{G, \mathcal{K}_{q+1}\}$. The figure is

$$\begin{array}{ccccc} A^q & \xrightarrow{\tau} & \text{Hom} \{\mathcal{B}_{q+1}, G\} & \xrightarrow{\eta} & \text{Ext} \{G, \mathcal{K}_{q+1}\} \\ \downarrow T & & \downarrow T_h^{**} & & \downarrow T_e^{**} \\ A'^q & \xrightarrow{\tau'} & \text{Hom} \{\mathcal{B}'_{q+1}, G\} & \xrightarrow{\eta'} & \text{Ext} \{G, \mathcal{K}'_{q+1}\} \end{array}$$

where T_h^{**} , T_e^{**} are again the induced homomorphisms. If $z^q \in A^q(K, G)$ is given, $\phi = \tau z^q$ is defined on each coboundary δd^q as $\phi(\delta d^q) = z^q \cdot d^q$, while $\phi' = T_h^{**} \phi$ is defined in turn as

$$\phi'(\delta d'^q) = \phi(T^* \delta d'^q) = \phi(\delta T^* d'^q) = z^q \cdot (T^* d'^q).$$

On the other hand, $\chi = \tau'(Tz^q)$ is defined on a coboundary $\delta d'^q$ of K' as

$$\chi(\delta d'^q) = (Tz^q) \cdot d'^q = z^q \cdot (T^* d'^q).$$

The results are identical, so $T_h^{**} \tau = \tau' T$. Now the "naturality" theorem for group extensions showed that T permutes with η , as in $T_e^{**} \eta = \eta' T_h^{**}$. Combination of these results gives

$$(\eta' \tau') T = T_e^{**} (\eta \tau).$$

This is the required commutativity condition, for $\eta \tau$ is the isomorphism envisaged in Theorem 31.3.

39. Čech's homology groups

We now briefly outline Čech's method of defining the homology and cohomology groups for a space X . Let U_α be a finite open covering of X and N_α the nerve of U_α . If U_β is a refinement of U_α we write $\alpha < \beta$. For $\alpha < \beta$ we have a chain transformation $T_{\alpha\beta}: N_\beta \rightarrow N_\alpha$ defined as follows: for each open set of the covering U_β select a set of U_α containing it; this maps the vertices of N_β into the vertices of N_α and leads to a simplicial mapping $T_{\alpha\beta}$. This chain transformation is not defined uniquely, but the induced homomorphisms

$$T_{\alpha\beta}: H^q(N_\beta, G) \rightarrow H^q(N_\alpha, G),$$

$$T_{\beta\alpha}^*: \mathcal{K}_q(N_\alpha, G) \rightarrow \mathcal{K}_q(N_\beta, G)$$

are unique. Using the directed system of all the finite open coverings of X we define²⁸

$$(39.1) \quad \mathcal{K}^q(X, G) = \varprojlim H^q(N_\alpha, G)$$

$$(39.2) \quad \mathcal{K}_q(X, G) = \varinjlim \mathcal{K}_q(N_\alpha, G).$$

In (39.2) the groups are all discrete. In (39.1) G can be any generalized topological group and $\mathcal{K}^q(X, G)$, as an inverse limit of generalized topological groups, also is a generalized topological group. If G has the property that each of its subgroups mG ($m = 2, 3, \dots$) is closed in G , the finiteness of each N_α implies that $H^q(N_\alpha, G)$ and hence $\mathcal{K}^q(X, G)$ is topological. If G does not have this property, it would still be possible to consider the group

$$\varprojlim H^q(N_\alpha, G) = \varprojlim [H^q(N_\alpha, G)/\bar{0}].$$

This group is always topological but its relation to the other groups is rather obscure.

In view of (37.6) the subgroups $Q^q(N_\alpha, G)$ of $H^q(N_\alpha, G)$ form an inverse system. We define

$$(39.3) \quad \mathcal{Q}^q(X, G) = \varprojlim Q^q(N_\alpha, G).$$

Clearly \mathcal{Q}^q is a subgroup of $\mathcal{K}^q(X, G)$.

Similarly, in view of (37.7), the subgroups $\mathcal{P}_q(N_\alpha, G)$ of $\mathcal{K}_q(N_\alpha, G)$ form a direct system so we define

$$(39.4) \quad \mathcal{P}_q(X, G) = \varinjlim \mathcal{P}_q(N_\alpha, G).$$

\mathcal{P}_q is a subgroup of $\mathcal{K}_q(X, G)$.

LEMMA 39.1. *The Kronecker index establishes a pairing of $\mathcal{K}^q(X, G)$ and $\mathcal{K}_q(X, I)$ with values in G ; under this pairing*

$$\mathcal{Q}^q(X, G) = \text{Annih } \mathcal{K}_q(X, I).$$

LEMMA 39.2. *Let G be discrete and $\hat{G} = \text{Char } G$. The Kronecker index establishes a dual pairing of $\mathcal{K}^q(X, \hat{G})$ and $\mathcal{K}_q(X, G)$ with values in the group P of reals mod 1; under this pairing*

$$\mathcal{K}_q(X, G) \cong \text{Char } \mathcal{K}^q(X, \hat{G})$$

$$\mathcal{P}_q(X, G) = \text{Annih } \mathcal{Q}^q(X, \hat{G}).$$

Both lemmas have been established for each of the complexes N_α . The passage to the limit is possible in view of formula (37.1.)

In $\mathcal{K}_q(X, G)$ we also consider the subgroup $\mathcal{J}_q(X, G)$ of all elements of finite

²⁸ For more detail see Lefschetz [7]. Although the definition of the homology and cohomology groups given here is valid for any space X , it is well known that its interest is restricted to compact spaces only. This is due to the fact that only in compact spaces is the family of finite open coverings cofinal with the family of all open coverings.

order. Since each approximating group $\mathcal{K}_q(N_\alpha, G)$ has a finite set of generators, one can show, by arguments resembling those of §24, that

$$\mathcal{I}_q(X, G) = \varinjlim \mathcal{I}_q(N_\alpha, G).$$

40. Formulas for a general space

Using the formulas for complexes and applying a straightforward passage to the limit we obtain here some relations for $\mathcal{K}^q(X, G)$ and $\mathcal{K}_q(X, G)$ in terms of the groups $\mathcal{K}_q(X, I)$ and $\mathcal{I}_{q+1}(X, I)$. The results are not as complete as in the case of a complex.

THEOREM 40.1. *For a space X and a generalized topological coefficient group G the subgroup \mathcal{Q}^q of the Čech homology group is expressible, in terms of a co-torsion group, as*

$$(40.1) \quad \mathcal{Q}^q(X, G) \cong \text{Ext}^* \{G, \mathcal{I}_{q+1}(X, I)\},$$

while the corresponding factor group $\mathcal{K}^q(X, G)/\mathcal{Q}^q(X, G)$ is isomorphic to a subgroup of $\text{Hom} \{\mathcal{K}_q(X, I), G\}$.

If G/mG is compact and topological for $m = 2, 3, \dots$ then

$$(40.2) \quad \mathcal{K}^q(X, G)/\mathcal{Q}^q(X, G) \cong \text{Hom} \{\mathcal{K}_q(X, I), G\}.$$

PROOF. For each nerve N_α we have (Theorem 32.1)

$$\mathcal{Q}^q(N_\alpha, G) \cong \text{Ext} \{G, \mathcal{I}_{q+1}(N_\alpha, I)\}$$

The groups on either side form inverse systems and it follows from Theorem 38.1 and Lemma 20.2 that the limits of these systems are isomorphic,

$$\mathcal{Q}^q(X, G) \cong \varinjlim \text{Ext} \{G, \mathcal{I}_{q+1}(N_\alpha, I)\}.$$

However since $\mathcal{I}_{q+1}(X, I) = \varinjlim \mathcal{I}_{q+1}(N_\alpha, I)$ and the groups $\mathcal{I}_{q+1}(N_\alpha, I)$ are finite it follows from Theorem 24.2 that the limit on the right is $\text{Ext}^* \{G, \mathcal{I}_{q+1}\}$. This proves formula (40.1).

From Theorem 32.1 we also have

$$H^q(N_\alpha, G)/Q^q(N_\alpha, G) \cong \text{Hom} \{\mathcal{K}_q(N_\alpha, I), G\},$$

and again the limits of the two inverse systems are isomorphic in view of Theorem 38.1. Consequently from Theorem 21.1 we get

$$\varinjlim [H^q(N_\alpha, G)/Q^q(N_\alpha, G)] \cong \text{Hom} \{\mathcal{K}_q(X, I), G\}.$$

Now it follows from (20.1) (Chap. IV) that the group

$$\mathcal{K}^q(X, G)/\mathcal{Q}^q(X, G) = \varinjlim H_\alpha^q / \varinjlim Q_\alpha^q$$

is isomorphic with a subgroup of the group $\varinjlim (H_\alpha^q/Q_\alpha^q)$. This proves the second assertion of the theorem. The subgroup will turn out to be the whole group whenever we are able to prove that $Q^q(N_\alpha, G)$ are compact topological groups.

Suppose now that G/mG is compact and topological for $m = 2, 3, \dots$

Given a cyclic group T of order $m \geq 2$ we have $\text{Ext } \{G, T\} \cong G/mG$ (Corollary 11.2) and consequently $\text{Ext } \{G, T\}$ is compact and topological. It follows that $\text{Ext } \{G, T\}$ is compact and topological for every finite group T . In particular the groups

$$\mathcal{Q}^q(N_\alpha, G) \cong \text{Ext } \{G, \mathcal{I}_{q+1}(N_\alpha, I)\}$$

are all compact and topological.

This completes the proof of the theorem. Notice that if G/mG is compact and topological for $m = 2, 3, \dots$ then the group $\mathcal{Q}^q(X, G)$, as a limit of compact topological groups, is compact and topological.

If G is discrete and has no elements of finite order, or if \mathcal{I}_{q+1} is countable, then by Theorem 24.4 and Corollary 24.1, the group Ext^* in (40.1) may be replaced by Ext/Ext_f . In particular if $G = I$ then by Theorems 17.1 and 40.1,

$$(40.3) \quad \mathcal{Q}^q(X, I) \cong \text{Char } \mathcal{I}_{q+1}(X, I),$$

$$(40.4) \quad \mathcal{K}^q(X, I)/\mathcal{Q}^q(X, I) \cong \text{Hom } \{\mathcal{K}_q(X, I), I\}.$$

THEOREM 40.2. *The Čech homology group $\mathcal{K}^q(X, G)$ of a space X over a compact topological group G has a subgroup \mathcal{Q}^q , with factor group $\mathcal{K}^q/\mathcal{Q}^q$, both expressible in terms of integral cohomology groups of X as*

$$(40.5) \quad \mathcal{Q}^q(X, G) \cong \text{Char Hom } \{G, \mathcal{I}_{q+1}(X, I)\},$$

$$(40.6) \quad \mathcal{K}^q(X, G)/\mathcal{Q}^q(X, G) \cong \text{Hom } \{\mathcal{K}_q(X, I), G\}.$$

PROOF. From Theorem 40.1 we have $\mathcal{Q}^q \cong \text{Ext}^* \{G, \mathcal{I}_{q+1}\}$. However since G is compact topological we have $\text{Ext}^* \{G, \mathcal{I}_{q+1}\} \cong \text{Ext } \{G, \mathcal{I}_{q+1}\}$ (Corollary 24.1) and $\text{Ext } \{G, \mathcal{I}_{q+1}\} \cong \text{Char Hom } \{G, \mathcal{I}_{q+1}\}$ (Theorem 15.1). This proves formula (40.5). We recall here that only continuous homomorphisms are considered. Formula (40.6) is a consequence of (40.2).

THEOREM 40.3. *The Čech cohomology groups $\mathcal{K}_q \supset \mathcal{P}_q$ of a space X over a discrete coefficient group G can be expressed, in part, in terms of the integral cohomology groups as*

$$(40.7) \quad \mathcal{P}_q(X, G) \cong G \circ \mathcal{K}_q(X, I),$$

$$(40.8) \quad \mathcal{K}_q(X, G)/\mathcal{P}_q(X, G) \cong \text{Hom } \{\text{Char } G, \mathcal{I}_{q+1}(X, I)\}.$$

PROOF. Let $\hat{G} = \text{Char } G$. Since $\mathcal{K}_q(X, G) \cong \text{Char } \mathcal{K}^q(X, \hat{G})$ and $\mathcal{P}_q = \text{Annih } \mathcal{Q}^q(X, \hat{G})$ we have

$$\mathcal{P}_q(X, G) \cong \text{Char } [\mathcal{K}^q(X, \hat{G})/\mathcal{Q}^q(X, \hat{G})],$$

and using Theorems 40.2 and 18.1 we get

$$\mathcal{P}_q(X, G) \cong \text{Char Hom } \{\mathcal{K}_q(X, I), \text{Char } G\} \cong G \circ \mathcal{K}_q(X, I).$$

This formula could have been proved directly, passing to the limit with $\mathcal{P}_q(N_\alpha, G) \cong G \circ \mathcal{K}_q(N_\alpha, I)$. Since also $\mathcal{K}_q/\mathcal{P}_q \cong \text{Char } \mathcal{Q}^q(X, \text{Char } G)$, formula (40.8) is a consequence of Theorem 40.2.

The theorems and proofs carry over without change to the homology theory of X modulo a closed subset. Another generalization can be obtained by replacing the space X by a net of complexes, as defined by Lefschetz ([7] Ch. VI).

We are unable to answer the question whether $\mathcal{Q}^q(X, G)$ and $\mathcal{P}_q(X, G)$ are direct factors of $\mathcal{K}^q(X, G)$ and $\mathcal{K}_q(X, G)$. This is why we do not obtain expressions for $\mathcal{K}^q(X, G)$ and $\mathcal{K}_q(X, G)$ in terms of $\mathcal{K}_q(X, I)$ and $\mathcal{J}_{q+1}(X, I)$. The best we achieve in the case of a general space X is a description of the subgroups \mathcal{Q}^q and \mathcal{P}_q and of the corresponding factor groups, leaving the direct product proposition undecided.²⁹

In the following sections of this chapter we shall discuss the case when X is a compact metric space, using the method of the fundamental complex. In this case we are able to obtain complete results, including the direct product decomposition.

41. The case $q = 0$

Before we proceed with the treatment of compact metric spaces we will discuss some details connected with the definition of the homology and cohomology groups for the dimension zero.

Let K be a finite simplicial complex. If we assume that there are no cells of dimension less than zero then every 0-chain will be a 0-cycle and the groups $H^0(K, G)$ and $\mathcal{K}_0(K, G)$ will be isomorphic to the product of G by itself n times, n being the number of components of K .

An alternate procedure is to consider K "augmented" by a single (-1) -cell σ^{-1} such that $[\sigma_i^0; \sigma^{-1}] = 1$ for all σ_i^0 . In this case, given a 0-chain $c^0 = \sum g_i \sigma_i^0$, we have $\partial c^0 = (\sum g_i) \sigma^{-1}$ and consequently c^0 is a cycle if and only if $\sum g_i = 0$. The cohomology group gets affected also because the cocycle $\sum \sigma_i^0$ that was not a coboundary in the first approach is a coboundary in the augmented complex, since $\delta \sigma^{-1} = \sum \sigma_i^0$. It turns out that $H^0(K, G)$ and $\mathcal{K}_0(K, G)$ are isomorphic to the product of G by itself $n - 1$ times.

In defining the groups $\mathcal{K}^0(X, G)$ and $\mathcal{K}_0(X, G)$ for a space we again have two alternatives according as the nerves N_α are augmented or not.

Both the augmented and unaugmented complexes are abstract complexes in the sense of Ch. V and therefore all our previous results hold for either definition of \mathcal{K}^0 and \mathcal{K}_0 . However in the discussion of compact metric spaces that follows there is an advantage in considering the nerves as augmented complexes, so as to have $\mathcal{K}^0(X, G) = \mathcal{K}_0(X, G) = 0$ if X is a connected space.

42. Fundamental complexes

Let X be a compact metric space. There is then a sequence U_n ($n = 0, 1, \dots$) of finite open coverings of X such that U_n is a refinement of U_{n-1} and every finite

²⁹ Steenrod [9] §10 brings an argument, which if correct would settle the question positively. Unfortunately an error occurs on p. 681, line 5. The error was noticed by C. Chevalley, who has also constructed an example showing that the argument could not be corrected in the general case. If X is metric compact, Steenrod's argument can be corrected to give the desired direct product decomposition (see §44 below).

open covering of X has some U_n as a refinement. This last property asserts that in the directed family of all the finite open coverings of X the sequence $\{U_n\}$ constitutes a cofinal subfamily and therefore the Čech homology and cohomology group can be equivalently defined using only the sequence of coverings U_n . We shall assume that U_0 is a covering consisting of only one set, namely X itself, so that the nerve N_0 of U_0 is a vertex. For each n we select a projection $T_n: N_n \rightarrow N_{n-1}$ of the nerve of U_n into the nerve of U_{n-1} . The projections $N_n \rightarrow N_{n-k}$ we define by transitivity.

We now define the fundamental complex K of X as follows. The complexes N_n for $n = 0, 1, \dots$ shall be disjoint subcomplexes of K . For each $n = 1, 2, \dots$ and each simplex σ^q of N_n we introduce a new $(q+1)$ -cell $\mathcal{D}\sigma^q$ whose boundary is $T_n\sigma^q - \sigma^q - \mathcal{D}\partial\sigma^q$. This formula gives a recursive definition of the incidence numbers.

In order to give a more intuitive picture of K we may consider each of the nerves N_n as a geometric simplicial complex, the projection T_n can then be regarded as a continuous simplicial transformation; that is, as linear on every simplex σ^q of N_n , while $\mathcal{D}\sigma^q$ can be visualized as a deformation prism consisting of intervals joining each point of σ^q with its image under T_n . With this interpretation K becomes a geometric complex and the cells $\mathcal{D}\sigma^q$ can be subdivided so as to furnish a simplicial subdivision of K . It is clear from this picture that K can be contracted to a point, namely by moving every point up its projection lines towards the vertex N_0 .

The complex K is countable and is locally finite; i.e., both closure and star finite. Viewing K as a closure finite complex, we can define finite cycles and infinite cocycles. However, since K is contractible all the homology group with finite cycles will vanish. Using the results of Ch. V we conclude that the cohomology groups with infinite cocycles also will vanish. Consequently, regarded as a closure finite complex, the structure of K is trivial. If we approach K as a star finite complex we obtain cohomology groups with finite cocycles and homology groups with infinite cycles. Regarded this way the complex K furnishes a true picture of the combinatorial structure of the space X .

43. Relations between a space and its fundamental complex

THEOREM 43.1. *The compact metric space X and its fundamental complex K are linked by isomorphisms*

$$(43.1) \quad \mathcal{K}^q(X, G) \cong H_w^{q+1}(K, G),$$

$$(43.2) \quad \mathcal{K}_q(X, G) \cong \mathcal{K}_{q+1}(K, G).$$

We shall restrict ourselves here to indicate the definitions of the isomorphisms without going into the complete proof, which involves lengthy but straightforward calculations.³⁰

Let z^q be an element of $\mathcal{K}^q(X, G)$. Then z^q can be represented by a sequence

³⁰ This proof is closely related to one given by Steenrod; see [10], §4.

of cycles $z_n^q \in Z^q(N_n, G)$ such that $z_{n-1}^q - T_n z_n^q \in B^q(N_{n-1}, G)$. For each $n = 1, 2, \dots$ select a chain c_{n-1}^{q+1} in N_{n-1} such that

$$\partial c_{n-1}^{q+1} = z_{n-1}^q - T_n z_n^q,$$

and consider the chain

$$z^{q+1} = \sum_{n=1}^{\infty} c_{n-1}^{q+1} + \sum_{n=1}^{\infty} \mathcal{D} z_n^q.$$

We verify that

$$\begin{aligned} \partial z^{q+1} &= \sum_{n=1}^{\infty} (z_{n-1}^q - T_n z_n^q) + \sum_{n=1}^{\infty} (T_n z_n^q - z_n^q - \mathcal{D} \partial z_n^q) \\ &= z_0^q - \sum_{n=1}^{\infty} \mathcal{D} \partial z_n^q = 0, \end{aligned}$$

since $\partial z_n^q = 0$, while $z_0^q = 0$ for $q \geq 0$, $z_0^0 = 0$ by §41. Consequently z^{q+1} is a cycle of K . If instead of $\{c_n^{q+1}\}$ we use a sequence $\{\bar{c}_n^{q+1}\}$ to define a cycle \bar{z}^{q+1} , then

$$z^{q+1} - \bar{z}^{q+1} = \sum_{n=1}^{\infty} (c_{n-1}^{q+1} - \bar{c}_{n-1}^{q+1})$$

Each term $c_{n-1}^{q+1} - \bar{c}_{n-1}^{q+1}$ is a finite cycle and therefore bounds in K , therefore $z^{q+1} - \bar{z}^{q+1}$ is a weakly bounding cycle and z^{q+1} determines uniquely an element $z^{q+1} \in H_w^{q+1}(K, G)$. We define

$$\phi(z^q) = z^{q+1}.$$

Now let $w^q \in \mathcal{H}_q(X, G)$. The element w^q can be represented for suitable n by a single cocycle $w^q \in \mathcal{Z}_q(N_n, G)$. We verify that $\mathcal{D}w^q$ is then a $(q+1)$ -cocycle of K . Using the formula

$$\delta w^q = \mathcal{D}T_n^* w^q - \mathcal{D}w^q \text{ in } K,$$

and the fact that \mathcal{D} and δ commute we show that $\mathcal{D}w^q$ determines uniquely an element w^{q+1} of $\mathcal{H}_q(K, G)$. We define

$$\psi(w^q) = w^{q+1}.$$

We also notice that the pair of isomorphisms ϕ, ψ preserves the Kronecker index

$$(43.3) \quad \phi(z^q) \cdot \psi(w^q) = z^q \cdot w^q.$$

If X_0 is a closed subset of X then every covering U_n of X determines a covering of X_0 whose nerve L_n is a subcomplex of the nerve N_n of U_n . The subcomplex

$$L = \sum_{n=1}^{\infty} L_n + \sum_{n=1}^{\infty} \mathcal{D}L_n$$

of K is then a fundamental complex of X_0 . The isomorphisms (43.1) and (43.2) of Theorem 43.1 can be generalized as follows

$$(43.1') \quad \mathcal{K}^q(X \bmod X_0, G) \cong H_w^{q+1}(K \bmod L, G)$$

$$(43.2') \quad \mathcal{K}_q(X - X_0, G) \cong \mathcal{K}_{q+1}(K - L, G).$$

44. Formulas for a compact metric space

Using the fundamental complex and the results of Ch. V we shall now establish theorems for a compact metric space quite analogous to the ones proved for a complex in Ch. V.

THEOREM 44.1. *The Čech homology groups of a compact metric space X over a generalized topological coefficient group G can be expressed in terms of the integral cohomology groups $\mathcal{K}_q = \mathcal{K}_q(X, I)$, $\mathcal{J}_{q+1} = \mathcal{J}_{q+1}(X, I)$ as*

$$\mathcal{K}^q(X, G) \cong \text{Hom} \{ \mathcal{K}_q, G \} \times (\text{Ext} \{ G, \mathcal{J}_{q+1} \} / \text{Ext}_f \{ G, \mathcal{J}_{q+1} \}).$$

More precisely, in terms of the subhomology group \mathcal{Q}^q of (39.3) we have

$$(44.1) \quad \mathcal{Q}^q(X, G) \text{ is a direct factor of } \mathcal{K}^q(X, G),$$

$$(44.2) \quad \mathcal{Q}^q(X, G) \cong \text{Ext} \{ G, \mathcal{J}_{q+1} \} / \text{Ext}_f \{ G, \mathcal{J}_{q+1} \},$$

$$(44.3) \quad \mathcal{K}^q(X, G) / \mathcal{Q}^q(X, G) \cong \text{Hom} \{ \mathcal{K}_q, G \}.$$

To prove the theorem we use the fact that the Kronecker intersection is preserved under the pair of isomorphisms ϕ, ψ of the previous section. Consequently, since

$$\mathcal{Q}^q(X, G) = \text{Annih } \mathcal{K}_q(X, I) \text{ in } \mathcal{K}^q(X, G),$$

$$Q_w^{q+1}(K, G) = \text{Annih } \mathcal{K}_{q+1}(K, I) \text{ in } H_w^{q+1}(K, G),$$

we have

$$\phi[\mathcal{Q}^q(X, G)] = Q_w^{q+1}(K, G),$$

and the theorem becomes a consequence of Theorems 43.1 and 32.2.

THEOREM 44.2. *The Čech cohomology groups of a compact metric space X with coefficients in a discrete group G can be expressed in terms of the integral cohomology groups $\mathcal{K}_q = \mathcal{K}_q(X, I)$, $\mathcal{J}_{q+1} = \mathcal{J}_{q+1}(X, I)$ as*

$$\mathcal{K}_q(X, G) \cong (G \circ \mathcal{K}_q) \times \text{Hom} \{ \text{Char } G, \mathcal{J}_{q+1} \}.$$

More precisely, in terms of the subgroup \mathcal{P}_q of (39.4), we have

$$(44.4) \quad \mathcal{P}_q(X, G) \text{ is a direct factor of } \mathcal{K}_q(X, G),$$

$$(44.5) \quad \mathcal{P}_q(X, G) \cong G \circ \mathcal{K}_q,$$

$$(44.6) \quad \mathcal{K}_q(X, G) / \mathcal{P}_q(X, G) \cong \text{Hom} \{ \text{Char } G, \mathcal{J}_{q+1} \}.$$

To prove the theorem we notice that

$$\begin{aligned}\mathcal{P}_q(X, G) &= \text{Annih } \mathcal{Q}^q(X, \text{Char } G) \text{ in } \mathcal{K}_q(X, G), \\ \mathcal{P}_{q+1}(K, G) &= \text{Annih } \mathcal{Q}_w^{q+1}(K, \text{Char } G) \text{ in } \mathcal{K}_{q+1}(K, G),\end{aligned}$$

and therefore

$$\psi[\mathcal{P}_q(X, G)] = \mathcal{P}_{q+1}(K, G)$$

and the theorem becomes a consequence of Theorems 43.1 and 33.1.

All these results remain valid for the homologies of X modulo a closed subset.

We now proceed to compare the results obtained here for the metric compact case with the results of §40 concerning general spaces.

Statements (44.1) and (44.4) contain a positive solution for the direct product problem which is still unsolved for the general space. Formula (44.3) was proved in (40.2) for general spaces only under the additional condition that G/mG be compact and topological for $m = 2, 3, \dots$. Formula (44.2) was proved for general spaces under the form

$$\mathcal{Q}^q(X, G) \cong \text{Ext}^* \{G, \mathcal{I}_{q+1}(X, I)\}$$

which is equivalent to (44.2) because

$$\text{Ext}^* \{G, T\} \cong \text{Ext} \{G, T\} / \text{Ext}_f \{G, T\}$$

for countable groups T with only elements of finite order (Theorem 24.4) and the group $\mathcal{I}_{q+1}(X, I) \cong \mathcal{I}_{q+2}(K, I)$ is countable for a compact metric X , since K is countable.

Formulas (44.5) and (44.6) coincide with the ones proved in Theorem 40.3 for a general space.

45. Regular cycles

Using the concept of a "regular cycle" Steenrod ([10]) has defined a new homology group $H^q(X, G)$ of "regular" cycles, for a compact metric space X . This group is useful especially in the case when X is a subset of the n -sphere S^n , because it provides information about the structure of the open set $S^n - X$.

Steenrod ([10], Theorem 7) has proved that if K denotes a fundamental complex of X then

$$(45.1) \quad H^q(X, G) \cong H^q(K, G).$$

From this, using Theorems 43.1 and 32.1 we derive the formula

$$(45.2) \quad H^q(X, G) \cong \text{Hom} \{ \mathcal{K}_{q-1}(X, I), G \} \times \text{Ext} \{ G, \mathcal{K}_q(X, I) \},$$

for $q > 0$. This formula expresses $H^q(X, G)$ in terms of $\mathcal{K}_{q-1}(X, I)$ and $\mathcal{K}_q(X, I)$ and hence shows that, essentially, $H^q(X, G)$ is no *new* invariant.

Let us specialize formula (45.2), assuming that $q = 1$, and that X is connected. We have then $\mathcal{K}_0(X, I) = 0$ and therefore

$$(45.3) \quad H^1(X, G) \cong \text{Ext} \{ G, \mathcal{K}_1(X, I) \}.$$

Let us further assume $G = I$ and that X is one of the solenoids Σ . Since Σ is a connected, compact abelian group we have $H^1(\Sigma, P) \cong \Sigma$ (Steenrod [9], Theorem 15) where P (Steenrod's \mathfrak{K}) is the group of reals mod 1. Further, since $\text{Char } I \cong P$ we have $H_1(\Sigma, I) \cong \text{Char } H^1(\Sigma, P) \cong \text{Char } \Sigma$. Hence finally

$$(45.4) \quad H^1(\Sigma, I) \cong \text{Ext} \{I, \text{Char } \Sigma\}.$$

This group will be explicitly computed in Appendix B; it was the starting point of this investigation (see introduction).

Steenrod has defined a subgroup $\tilde{H}^q(X, G)$ of $H^q(X, G)$ by considering regular cycles that are sums of finite cycles. He has also proved that under the isomorphism (45.1) this group is mapped onto the subgroup $B_w^q(K, G)/B^q(K, G)$ of $H^q(K, G)$.

We shall now show that, for $q > 1$,

$$(45.5) \quad \tilde{H}^q(X, G) \cong \text{Ext}_f \{G, \mathfrak{K}_q(X, I)\}.$$

$$(45.6) \quad H^q(X, G)/\tilde{H}^q(X, G) \cong \mathfrak{K}^{q-1}(X, G).$$

In fact, from Theorems 31.3 and 43.1 we deduce that $B_w^q(K, G)/B^q(K, G) \cong \text{Ext}_f \{G, \mathfrak{K}_{q+1}(K, I)\} \cong \text{Ext}_f \{G, \mathfrak{K}_q(X, I)\}$. This proves (45.5). In order to prove (45.6) notice that $H^q(X, G)/\tilde{H}^q(X, G) \cong H^q(K, G)/[B_w^q(K, G)/B^q(K, G)] \cong H_w^q(K, G) \cong \mathfrak{K}^{q-1}(X, G)$.

Formulas (45.5) and (45.6) provide a splitting of $H^q(X, G)$ different from the one used in (45.2). The isomorphism (45.6) was established by Steenrod [10], who has also shown that \tilde{H}^q can be computed using G and $\mathfrak{K}_q(X, I)$, without however getting the explicit formula (45.5).

From (45.5) we immediately deduce the theorem of Steenrod that $\tilde{H}^q(X, G) = 0$ and $H^q(X, G) \cong \mathfrak{K}^{q-1}(X, G)$ whenever $\mathfrak{K}_q(X, I)$ has a finite number of generators.

APPENDIX A. COEFFICIENT GROUPS WITH OPERATORS

In many topological investigations it is convenient to construct homology groups $H^q(K, G)$ in cases when G is not just a group, but a ring or even a field. More generally, G can be allowed to be a group with operators. We show here that our results extend unchanged to such cases, and in particular, that the resulting homology groups are still completely determined by the integral cohomology groups.

G is called a *group with operators* Ω if G is a generalized topological group, Ω a space, and if to each element $\omega \in \Omega$ and each $g \in G$ there is assigned an element $\omega g \in G$ (the result of operating on g with ω), in such wise that

$$(i) \quad \omega g \text{ is a continuous function of the pair } (\omega, g),$$

$$(ii) \quad \omega(g_1 + g_2) = \omega g_1 + \omega g_2 \quad (g_1, g_2 \in G).$$

It then follows that each element ω determines a (continuous) homomorphism $g \rightarrow \omega g$ of G into G ; however, distinct elements of Ω need not determine distinct

homomorphisms. The set Ω may have a discrete topology, or may even consist of just one operator ω .

If both G_1 and G_2 have operators Ω , a homomorphism (or isomorphism) ϕ of G_1 into G_2 is said to be Ω -allowable if $\phi[\omega g_1] = \omega[\phi g_1]$ for all $g_1 \in G_1$, $\omega \in \Omega$.

If G has operators Ω , a subgroup $S \subset G$ is said to be allowable if $\omega(S) \subset S$ for all $\omega \in \Omega$. The operators Ω may then be applied in natural fashion to the factor group G/S , by setting $\omega(g + S) = \omega g + S$. Then G/S is a group with operators Ω , and the natural homomorphism of G on G/S is allowable.³¹

If G is a group with operators Ω , the various groups introduced as functions of G in Chapters I–IV are also groups with operators. Specifically, let H be a discrete group, and for each $\theta \in \text{Hom } \{H, G\}$ define $\omega\theta$ as $[\omega\theta](h) = \omega[\theta(h)]$. Then $\omega\theta \in \text{Hom } \{H, G\}$, and

$$(A.1) \quad \text{Hom } \{H, G\} \text{ has operators } \Omega.$$

Furthermore, if $H = F/R$, where F is free, the groups $\text{Hom } \{F | R, G\}$ and $\text{Hom}_r \{R, G; F\}$ are allowable subgroups of $\text{Hom } \{R, G\}$, so

$$(A.2) \quad \text{Hom } \{R, G\} / \text{Hom } \{F | R, G\} \text{ has operators } \Omega.$$

Again, let f be a factor set of H in G , and define another factor set ωf by taking $[\omega f](h, k)$ as $\omega[f(h, k)]$. Then Ω becomes a space of operators for the group $\text{Fact } \{G, H\}$. Furthermore $\text{Trans } \{G, H\}$ is an allowable subgroup; therefore

$$(A.3) \quad \text{Ext } \{G, H\} \text{ has operators } \Omega.$$

In similar fashion one concludes that $\text{Ext}_r \{G, H\}$ and Ext/Ext_r have operators Ω .

As another case, take $\phi \in \text{Hom } \{G, H\}$ and define a homomorphism $\omega\phi \in \text{Hom } \{G, H\}$ by setting $[\omega\phi](g) = \phi[\omega(g)]$ for each $g \in G$. If G is compact or discrete, one may show that $\omega\phi$ is a continuous function of ω and ϕ . In this case, and for any generalized topological group H ,

$$(A.4) \quad \text{Hom } \{G, H\} \text{ has operators } \Omega.$$

In particular, if G is discrete or compact,

$$(A.5) \quad \text{Char } G \text{ has operators } \Omega.$$

Given these interpretations of all our basic groups as groups with operators, we next demonstrate that the various isomorphisms between these groups, as established in Chapters II–IV, are allowable. In particular, an inspection of the construction used to establish the fundamental Theorem 10.1 of Chapter II proves

$$(A.6) \quad \text{The isomorphism}$$

$$\text{Ext } \{G, H\} \cong \text{Hom } \{R, G\} / \text{Hom } \{F | R, G\},$$

where $H = F/R$, F free, is allowable.

³¹ Practically all the elementary formal facts about groups and homomorphisms apply to operator groups and allowable homomorphisms.

The same conclusion holds for the other isomorphisms stated in that theorem. Also, the isomorphism $\text{Ext } \{G, H\} \cong \text{Char Hom } \{G, H\}$ established in Theorem 15.1 for compact topological G and discrete H is allowable. The proof of this fact depends essentially on showing that the "trace" used in that theorem has the commutation property,

$$t(\omega\theta, \phi) = t(\theta, \omega\phi), \text{ for any } \theta \in \text{Hom } \{R, G\}, \text{ and } \phi \in \text{Hom } \{G, H\}.$$

The allowability of the other isomorphisms in Chapters II–IV is similarly established. The proofs are closely analogous to the "naturality" proofs of §12, except that here the operators apply to G , while in §12 the operator T applied to H .

Now turn to the homology groups. Let c^q be a chain in the star finite complex K , with coefficients chosen in the group G with operators Ω . For each $\omega \in \Omega$, define

$$\omega(c^q) = \omega(\sum_i g_i \sigma_i^q) = \sum_i (\omega g_i) \sigma_i^q;$$

since the result is a chain, and since the requisite continuity holds, the group $C^q(K, G)$ of q -chains has operators Ω . Moreover, $\omega\partial = \partial\omega$, so that both $Z^q(K, G)$ and $B^q(K, G)$ are allowable subgroups of C^q . Therefore

$$(A.7) \quad H^q(K, G) \text{ has operators } \Omega.$$

The essential tool in establishing the isomorphisms of Chap. V is the Kronecker index $c^q \cdot d^q$ for $d^q \in \mathcal{C}_q(K, I)$, $c^q \in C^q(K, G)$. We verify at once that

$$(A.8) \quad \omega(c^q \cdot d^q) = (\omega c^q) \cdot d^q \quad (\text{all } \omega \in \Omega).$$

Since the subgroup A^q of Z^q was defined as a certain annihilator under this Kronecker index (see (29.9)), it follows at once that A^q is an allowable subgroup of Z^q . Furthermore, the proof that A^q is a direct factor of C^q depended on a decomposition of \mathcal{C}_q as a direct product $\mathcal{C}_q = \mathcal{Z}_q \times \mathcal{D}_q$, for a suitably chosen group \mathcal{D}_q . In the notation of Lemma 16.2, we then had, by means of the Kronecker index (see the proof of Theorem 30.3)

$$C^q \cong \text{Hom } \{\mathcal{C}_q, G\} \cong \text{Hom } \{\mathcal{C}_q, G; \mathcal{D}_q, 0\} \times \text{Hom } \{\mathcal{C}_q, G; \mathcal{Z}_q, 0\}.$$

On the right both factors are allowable subgroups, and the isomorphism to the direct product is allowable;³² furthermore, the second factor is the one which corresponds to the subgroup A^q of C^q . Therefore C^q has a representation of the form $C^q = A^q \times D^q$, where D^q is an allowable subgroup, complementary to A^q . A similar decomposition holds for Z^q and thus for its factor group $H^q = Z^q/B^q$. In terms of the homology subgroup $Q^q = A^q/B^q$ determined by A^q , this proves

$$(A.9) \quad \text{The isomorphism } H^q \cong (H^q/Q^q) \times Q^q \text{ is allowable.}$$

The further analysis of these two factors, as carried out in Chapter V, all depended on the Kronecker index. In view of the property (A.8) of this index,

³² If A and B are two groups with operators Ω the direct product $A \times B$ has operators Ω defined by $\omega(a, b) = (\omega(a), \omega(b))$ for $\omega \in \Omega$.

and the property (A.6) of the basic group-extension theorem, we have

(A.10) *The isomorphisms*

$$H^q(K, G)/Q^q(K, G) \cong \text{Hom} \{ \mathcal{K}_q, G \},$$

$$Q^q(K, G) \cong \text{Ext} \{ G, \mathcal{K}_{q+1} \}$$

are allowable, as is the isomorphism $H^q \cong \text{Hom} \times \text{Ext}$, obtained by combining (A.9) and (A.10).

Similar remarks apply to the representation of the "weak" homology group H_w^q (Theorem 32.2), which is a factor group of H^q by an allowable subgroup. The same holds for the topologized homology group H_i^q (i.e., the isomorphisms of Theorem 34.2 are allowable), for in any topological group G with operators Ω , the continuity of the operators insures that the subgroup $\bar{0} \subset G$ is allowable (recall that $H_i^q = H^q/\bar{0}$).

Turn next to the analysis of the cohomology groups. The groups $\mathcal{C}_q(K, G)$, with G discrete, again have operators in Ω , under the natural definition. As in the case of the homology groups, we have

(A.11) $\mathcal{K}_q(K, G)$ has operators Ω .

The representation of these groups depended on duality; i.e., on the Kronecker index $c^q \cdot d^q$, for $c^q \in Z^q(K, \text{Char } G)$, $d^q \in Z^q(K, G)$. Given the various definitions of the effect of an operator ω , one shows easily that

$$(\omega c^q) \cdot d^q = c^q \cdot (\omega d^q) \quad (\text{all } \omega \in \Omega).$$

From this formula one may deduce that the well known isomorphism $\mathcal{K}_q(K, G) \cong \text{Char } H^q(K, \text{Char } G)$ is allowable. Thence it follows that the isomorphisms of Theorem 33.1 representing \mathcal{K}_q are allowable.

These considerations yield the following

ADDENDUM TO THE UNIVERSAL COEFFICIENT THEOREM. *If K is any star finite complex, G a group with operators Ω , then the homology groups of K (and, if G is discrete, the finite cohomology groups) with coefficients in G all have operators Ω . All these groups with their operators are determined by the group G (with its operators) and the cohomology groups of the finite integral cocycles of K .*

A similar discussion applies to the results of Chap. VI.

In many important cases the operators form a ring (or even a field). Let us assume then that Ω is a generalized topological ring; that is, a ring which is a generalized topological group under addition and in which the multiplication is continuous. Then G is called an Ω -modulus if G is a generalized topological group with operators Ω (i.e., conditions (i) and (ii) above hold) such that

$$(iii) \quad (\omega_1 \omega_2)g = \omega_1(\omega_2 g). \quad (\text{for } \omega_i \in \Omega, g \in G),$$

$$(iv) \quad (\omega_1 + \omega_2)g = \omega_1 g + \omega_2 g \quad (\text{for } \omega_i \in \Omega, g \in G).$$

In other words, addition and multiplication of operators are determined in the natural fashion from G .

If the standard coefficient group G is now assumed to be an Ω -modulus, simple

arguments will show that all the groups with operators Ω as described above are in fact Ω -moduli. Since the basic isomorphisms are still Ω -allowable, we conclude that the addendum to the universal coefficient group theorem still holds in these circumstances.

It is sometimes convenient to use a set Ω of operators in which only the addition or only the multiplication of operators is defined. More generally, we may consider a space Ω in which only certain sums $\omega_1 + \omega_2$ and products $\omega_1\omega_2$ are defined (and continuous); we then require that conditions (iii) and (iv) above hold only when the terms $\omega_1\omega_2$ or $\omega_1 + \omega_2$ are defined. The derived groups satisfy similar assumptions, and the universal coefficient theorem still holds.

If the coefficient group G is locally compact, one can always take the operators to form a ring, for any such group G has its endomorphism ring Ω_G as a natural ring of operators. Specifically, Ω_G is the additive group $\text{Hom } \{G, G\}$ of endomorphisms of G , with its usual topology (§3), and the multiplication $\omega_1\omega_2$ of two endomorphisms is defined by (iii) above. The requisite continuity properties of $\omega_1\omega_2$ and ωg are readily established, in virtue of the local compactness of G . Furthermore, if Ω is any other space of operators on G , each $\omega \in \Omega$ determines uniquely an endomorphism $\bar{\omega} \in \Omega_G$ with $\bar{\omega}g = \omega g$ for each g . The correspondence $\omega \rightarrow \bar{\omega}$ is a continuous mapping of Ω into Ω_G which preserves whatever sums and products may be present in Ω (assumed to satisfy (iii) and (iv)). Thus, any group, derived from G , which is an Ω_G -modulus is also a group with operators Ω , and any isomorphism between groups which is Ω_G -allowable is Ω -allowable. This indicates, that, for locally compact groups, one may restrict attention to operators of the ring Ω_G .

The most useful case is that in which the coefficient group is a field F , which is its own ring of operators. In this case all the homology groups, groups of homomorphisms, etc., become F -moduli; that is, vector spaces over F .

All these remarks suggest the following rather negative conclusion: *although in many applications it is convenient to consider a homology theory over coefficients which form more than merely a group, no new topological invariants can be so obtained.*

APPENDIX B. SOLENOIDS

Here we compute the one-dimensional homology group $H^1(\Sigma, I)$ of regular cycles for the solenoid³³ Σ , or the isomorphic group $\text{Ext } \{I, \text{Char } \Sigma\}$ (see (45.4)).

A solenoid is uniquely determined by a Steinitz G -number; that is, by a formal (infinite) product $G = \prod p_i^{e_i}$ of distinct primes with exponents e_i which are non-negative integers or ∞ . Any such number G can be represented (in many ways) as a formal product $G = a_1 a_2 \cdots a_n \cdots$ of ordinary integers a_i ; if G is not an ordinary integer, we can take each $a_i \geq 2$. Given such a representation of G , take replicas P_n of the additive group P of real numbers modulo 1, and let ϕ_n be the homomorphism which wraps P_n a_{n-1} times around P_{n-1} . The

³³ Solenoids were studied by L. Vietoris, *Math. Annalen* 97 (1927), p. 459, and more in detail by D. van Dantzig, *Fundam. Math.* 15 (1930), pp. 102–135. See also L. Pontrjagin [8], p. 171.

P_n then form an inverse system of groups, relative to the homomorphisms $\psi_{n+m,n} = \phi_{n+1} \cdots \phi_{n+m}$, and the solenoid Σ_G is defined as the limit $\Sigma_G = \varprojlim P_n$. Therefore $\text{Char } \Sigma_G = \varprojlim \text{Char } P_n$, where the groups form a direct system under the dual correspondences ϕ_n^* . Here $\text{Char } P_n$ is an isomorphic replica I_n of the additive group of integers, and ϕ_n^* maps I_n into I_{n+1} by multiplying each $x \in I_n$ by a_n . Therefore $\text{Char } \Sigma_G = \varprojlim I_n$ is a subgroup N_G of the additive group of rational numbers, consisting of all rationals of the form a/d_n , with a an integer and $d_n = a_1 \cdots a_{n-1}$. Alternatively, N_G consists of all rationals r/s with s a "divisor" of G ; hence N_G and Σ_G are uniquely determined by G , and are independent of the representation $G = a_1 a_2 \cdots a_n \cdots$.

A Steinitz G -number which is not an ordinary integer also determines a certain topological ring. Set $G = a_1 a_2 \cdots a_n \cdots$, $d_n = a_1 \cdots a_{n-1}$. In the ring I of integers, introduce as neighborhoods of zero the sets (d_n) of all multiples of d_n . Since the intersection of all these (d_n) is the zero element of I , these neighborhoods make I a topological ring. It can be embedded in a unique fashion in a minimal complete topological ring $I_G \supset I$, so that every element of I_G is a limit of a sequence of integers, under the given topology. This is one of the b -adic rings introduced by D. van Dantzig.³⁴ The additive group of I_G can be alternatively described as a limit of an inverse sequence; specifically, the factor group $I/(d_{n+1})$ has a natural homomorphism into $I/(d_n)$, and the limit group is $I_G \cong \varprojlim I/(d_n)$. In the special case when $G = p^\infty$ is an infinite power of a prime p , I_G is the ordinary ring of p -adic integers.

THEOREM. *If G is any Steinitz G -number which is not an ordinary integer, Σ_G the corresponding solenoid, and I_G the corresponding complete ring containing the ring I of integers, then*

$$(B.1) \quad \text{Ext } \{I, \text{Char } \Sigma_G\} \cong I_G/I.$$

PROOF. As above, $\text{Char } \Sigma_G$ is a group N_G of rationals, generated by the numbers $r_n = 1/d_n$ with relations $a_n r_{n+1} = r_n$. Therefore N_G can be represented as F/R , where F is a free group with generators z_1, z_2, \dots , and R the subgroup with generators $y_n = a_n z_{n+1} - z_n$, $n = 1, 2, \dots$. By the fundamental theorem on group extensions

$$(B.2) \quad \text{Ext } \{I, \text{Char } \Sigma_G\} \cong \text{Hom } \{R, I\} / \text{Hom } \{F | R, I\}.$$

Let $\theta \in \text{Hom } \{R, I\}$ and set

$$x(\theta) = \lim_{n \rightarrow \infty} [\theta y_1 + d_2 \theta y_2 + \cdots + d_n \theta y_n].$$

Then $x(\theta)$ is a well-defined element of I_G , and $\theta \rightarrow x(\theta)$ is a homomorphic mapping of $\text{Hom } \{R, I\}$ into I_G and thus, derivatively, into I_G/I . We assert that the kernel of the latter mapping is $\text{Hom } \{F | R, I\}$.

Assume first that $\theta \in \text{Hom } \{F | R, I\}$, and let θ^* be an extension of θ to F . Then

$$\theta(y_1 + d_2 y_2 + \cdots + d_n y_n) = -\theta^* z_1 + d_{n+1} \theta^* z_{n+1},$$

³⁴ Math. Annalen 107 (1932), pp. 587-626; Compositio Math. 2 (1935), pp. 201-223.

so that the limit $x(\theta)$ is $-\theta^*z_1$, which is an integer in I . Conversely, suppose that $x(\theta) \in I$, and set $x(\theta) = -c_1$. We then have

$$\theta y_1 + d_2 \theta y_2 + \cdots + d_n \theta y_n \equiv -c_1 \pmod{d_{n+1}}.$$

By successive applications of this condition we find integers c_n with $\theta y_n = a_n c_{n+1} - c_n^*$. The homomorphism $\theta^*z_n = c_n$ then provides an extension of θ to F , so that $\theta \in \text{Hom } \{F | R, I\}$.

Every element in I_θ is a limit of integers, hence has the form $\text{Lim } [b_1 + d_2 b_2 + \cdots + d_n b_n]$; therefore $\theta \rightarrow x(\theta)$ is a mapping onto I_θ . We thus have

$$(B.3) \quad \text{Hom } \{R, I\} / \text{Hom } \{F | R, I\} \cong I_\theta / I.$$

The correspondence is topological, as one may readily verify that both (generalized topological) groups carry the trivial topology in which the only open sets are zero and the whole space. Thus (B.2) and (B.3) prove the isomorphism (B.1).

By cardinal number considerations, one shows that the group I_θ/I is uncountable, hence not void. The formula (B.1) gives at once all the special properties of the homology group of the solenoid, as found by Steenrod [10] in his partial determination of this group.

THE UNIVERSITY OF MICHIGAN
HARVARD UNIVERSITY

BIBLIOGRAPHY

- [1] ALEXANDROFF, P. AND HOPF, H. *Topologie*, vol. I, Berlin, 1935.
- [2] BAER, R. *Erweiterungen von Gruppen und ihren Isomorphismen*, Math. Zeit. 38 (1934), pp. 375-416.
- [3] ČECH, E. *Les groupes de Betti d'un complexe infini*, Fund. Math. 25 (1935), pp. 33-44.
- [4] EILENBERG, S. *Cohomology and continuous mappings*, Ann. of Math. 41 (1940), pp. 231-251.
- [5] EILENBERG, S. AND MAC LANE, S. *Infinite Cycles and Homologies*, Proc. Nat. Acad. Sci. U. S. A. 27 (1941), pp. 535-539.
- [6] HALL, M. *Group Rings and Extensions, I*, Ann. of Math. 39 (1938), pp. 220-234.
- [7] LEFSCHETZ, S. *Algebraic Topology*, Amer. Math. Soc. Colloquium Series, vol. 27, New York, 1942.
- [8] PONTRJAGIN, L. *Topological Groups*, Princeton, 1939.
- [9] STEENROD, N. E. *Universal Homology Groups*, Amer. Journ. of Math. 58 (1936), pp. 661-701.
- [10] ———. *Regular Cycles of Compact Metric Spaces*, Ann. of Math. 41 (1940), pp. 833-851.
- [11] TURING, A. M. *The Extensions of a Group*, Compositio Math. 5 (1938), pp. 357-367.
- [12] WEIL, A. *L'Integration dans les groupes topologiques et ses applications*, Actualites Sci. et Ind. No. 869, Paris, 1940.
- [13] WHITNEY, H. *Tensor Products of Abelian Groups*. Duke Math. Journ. 4 (1938), pp. 495-528.
- [14] ———. *On matrices of integers and combinatorial topology*, Duke Math. Journ. 3 (1937), pp. 35-45.
- [15] ZASSENHAUS, H. *Lehrbuch der Gruppentheorie*, Hamburg. Math. Einzelschriften, 21, Leipzig, 1937.

INDEX

ALBERT, A. A. Quadratic forms permitting composition.....	161
ALBERT, A. A. Non-associative algebras. I. Fundamental concepts and isotopy.....	685
ALBERT, A. A. Non-associative algebras. II. New simple algebras.....	708
ARONSZAJN, N. Le correspondant topologique de l'unicité dans la théorie des équations différentielles.....	730
BIRKHOFF, G. Lattice-ordered groups.....	298
BOCHNER, S. On a theorem of Tannaka and Krein.....	56
BOCHNER, S., AND PHILLIPS, R. S. Absolutely convergent Fourier expan- sions for non-commutative normed rings.....	409
BLACKWELL, D. Idempotent Markoff chains.....	560
CHERN, SHIING-SHEN. On integral geometry in Klein spaces.....	178
CHERN, SHIING-SHEN. The geometry of isotropic surfaces.....	545
DOOB, J. L. The Brownian movement and stochastic equations.....	351
EILENBERG, S. Banach space methods in topology.....	568
EILENBERG, S., AND MACLANE, S. Group extensions and homology.....	757
ERDÖS, P. On the uniform distribution of the roots of certain polynomials.....	59
ERDÖS, P. On the asymptotic density of the sum of two sequences.....	65
ERDÖS, P. On the law of the iterated logarithm.....	419
ERDÖS, P. On an elementary proof of some asymptotic formulas in the theory of partitions.....	437
ERDÖS, P., AND SZEGÖ, G. On a problem of I. Schur.....	451
FREUDENTHAL, H. Neuaufbau der Endentheorie.....	261
FREUDENTHAL, H. Simplicialzerlegungen von beschränkter Flachheit....	580
FUBINI, G. On Abel's converse theorem.....	471
GARABEDIAN, H. L. Hausdorff integral transformations.....	501
GROVE, V. G. The transformation T of congruences.....	623
HADAMARD, J. The problem of diffusion of waves.....	510
HALMOS, P., AND VON NEUMANN, J. Operator methods in classical me- chanics II.....	332
HEINS, M. H. On the continuation of a Riemann surface.....	280
HOOKE, R. Linear p-adic groups and their Lie algebras.....	641
KAKUTANI, S. A proof that there exists a circumscribing cube around any bounded closed convex set in R^3	739
KAKUTANI, S. An extremum problem in product measure.....	742
KOLCHIN, E. R. Extensions of differential fields.....	724
MACKEY, G. W. Isomorphisms of normed linear spaces.....	244
MACLANE, S., AND EILENBERG, S. Group extensions and homology.....	757
MANN, H. B. A proof of the fundamental theorem on the density of sums of sets of positive integers.....	523
MANNING, RHODA. On the derivatives of the sections of bounded power series.....	617

MAYER, W. A new homology theory.....	370
MAYER, W. A new homology theory II.....	594
NESBITT, C. J., AND THRALL, R. M. On the modular representation of the symmetric group.....	656
NEUMANN, J. V., AND HALMOS, P. Operator methods in classical mechanics II.....	332
NEWMAN, M. H. A. On theories with a combinatorial definition of "equivalence".....	223
PHILLIPS, R. S. AND BOCHNER, S. Absolutely convergent Fourier expansions for non-commutative normed rings.....	409
SCOTT, W. M. On matrix algebras over an algebraically closed field.....	147
SHIFFMAN, M. Unstable minimal surfaces with several boundaries.....	197
SIEGEL, C. L. Iteration of analytic functions.....	607
SIEGEL, C. L. Note on automorphic functions of several variables.....	613
SMILEY, M. F. A remark on S. Kakutani's characterization of (L) spaces.....	528
STEENROD, N. E. Topological methods for the construction of tensor functions.....	116
SZÁSZ, O. Some new summability methods with applications.....	69
SZEGÖ, G., AND ERDÖS, P. On a problem of I. Schur.....	451
THRALL, R. M. On the decomposition of modular tensors (I).....	671
THRALL, R. M., AND NESBITT, C. J. On the modular representation of the symmetric group.....	656
TRJITZINSKY, W. J. Analytic theory of singular elliptic partial differential equation.....	1
WARD, M. The closure operators of a lattice.....	191
WEYL, H. On the differential equations of the simplest boundary-layer problems.....	381
WHITEHEAD, G. W. Homotopy properties of the real orthogonal groups.....	132
WHITEHEAD, G. W. On the homotopy groups of spheres and rotation groups.....	634
WHITMAN, P. M. Free lattices II.....	104
YOUNG, L. C. Generalized surfaces in the calculus of variations. I. Generalized Lipschitzian surfaces.....	84
YOUNG, L. C. Generalized surfaces in the calculus of variations. II.....	530
ZARISKI, O. A simplified proof for the resolution of singularities of an algebraic surface.....	583

INDEX

ALBERT, A. A. Quadratic forms permitting composition.....	161
ALBERT, A. A. Non-associative algebras. I. Fundamental concepts and isotopy.....	685
ALBERT, A. A. Non-associative algebras. II. New simple algebras.....	708
ARONSZAJN, N. Le correspondant topologique de l'unicité dans la théorie des équations différentielles.....	730
BIRKHOFF, G. Lattice-ordered groups.....	298
BOCHNER, S. On a theorem of Tannaka and Krein.....	56
BOCHNER, S., AND PHILLIPS, R. S. Absolutely convergent Fourier expan- sions for non-commutative normed rings.....	409
BLACKWELL, D. Idempotent Markoff chains.....	560
CHERN, SHIING-SHEN. On integral geometry in Klein spaces.....	178
CHERN, SHIING-SHEN. The geometry of isotropic surfaces.....	545
DOOB, J. L. The Brownian movement and stochastic equations.....	351
EILENBERG, S. Banach space methods in topology.....	568
EILENBERG, S., AND MACLANE, S. Group extensions and homology.....	757
ERDÖS, P. On the uniform distribution of the roots of certain polynomials.....	59
ERDÖS, P. On the asymptotic density of the sum of two sequences.....	65
ERDÖS, P. On the law of the iterated logarithm.....	419
ERDÖS, P. On an elementary proof of some asymptotic formulas in the theory of partitions.....	437
ERDÖS, P., AND SZEGÖ, G. On a problem of I. Schur.....	451
FREUDENTHAL, H. Neuaufbau der Endentheorie.....	261
FREUDENTHAL, H. Simplicialzerlegungen von beschränkter Flachheit....	580
FUBINI, G. On Abel's converse theorem.....	471
GARABEDIAN, H. L. Hausdorff integral transformations.....	501
GROVE, V. G. The transformation T of congruences.....	623
HADAMARD, J. The problem of diffusion of waves.....	510
HALMOS, P., AND VON NEUMANN, J. Operator methods in classical me- chanics II.....	332
HEINS, M. H. On the continuation of a Riemann surface.....	280
HOOKE, R. Linear p-adic groups and their Lie algebras.....	641
KAKUTANI, S. A proof that there exists a circumscribing cube around any bounded closed convex set in R^3	739
KAKUTANI, S. An extremum problem in product measure.....	742
KOLCHIN, E. R. Extensions of differential fields.....	724
MACKEY, G. W. Isomorphisms of normed linear spaces.....	244
MACLANE, S., AND EILENBERG, S. Group extensions and homology.....	757
MANN, H. B. A proof of the fundamental theorem on the density of sums of sets of positive integers.....	523
MANNING, RHODA. On the derivatives of the sections of bounded power series.....	617

MAYER, W. A new homology theory.....	370
MAYER, W. A new homology theory II.....	594
NESBITT, C. J., AND THRALL, R. M. On the modular representation of the symmetric group.....	656
NEUMANN, J. V., AND HALMOS, P. Operator methods in classical mechanics II.....	332
NEWMAN, M. H. A. On theories with a combinatorial definition of "equivalence".....	223
PHILLIPS, R. S. AND BOCHNER, S. Absolutely convergent Fourier expansions for non-commutative normed rings.....	409
SCOTT, W. M. On matrix algebras over an algebraically closed field.....	147
SHIFFMAN, M. Unstable minimal surfaces with several boundaries.....	197
SIEGEL, C. L. Iteration of analytic functions.....	607
SIEGEL, C. L. Note on automorphic functions of several variables.....	613
SMILEY, M. F. A remark on S. Kakutani's characterization of (L) spaces.....	528
STEENROD, N. E. Topological methods for the construction of tensor functions.....	116
SZÁSZ, O. Some new summability methods with applications.....	69
SZEGÖ, G., AND ERDÖS, P. On a problem of I. Schur.....	451
THRALL, R. M. On the decomposition of modular tensors (I).....	671
THRALL, R. M., AND NESBITT, C. J. On the modular representation of the symmetric group.....	656
TRJITZINSKY, W. J. Analytic theory of singular elliptic partial differential equation.....	1
WARD, M. The closure operators of a lattice.....	191
WEYL, H. On the differential equations of the simplest boundary-layer problems.....	381
WHITEHEAD, G. W. Homotopy properties of the real orthogonal groups.....	132
WHITEHEAD, G. W. On the homotopy groups of spheres and rotation groups.....	634
WHITMAN, P. M. Free lattices II.....	104
YOUNG, L. C. Generalized surfaces in the calculus of variations. I. Generalized Lipschitzian surfaces.....	84
YOUNG, L. C. Generalized surfaces in the calculus of variations. II.....	530
ZARISKI, O. A simplified proof for the resolution of singularities of an algebraic surface.....	583

L.A.R.1. 75.

INDIAN AGRICULTURAL RESEARCH
INSTITUTE LIBRARY, NEW DELHI.

[illegible]

I.A.R.I.—29-4-5—15,000